Contents through Networks

Anastasios Giovanidis

Sorbonne Université

Habilitation to Supervise Research (HdR) Thesis

Contents through Networks:
wireless access, caching, social diffusion

Authored by

Anastasios GIOVANIDIS

prepared at Sorbonne Université, LIP6 - Team NPA

Submitted in total fulfilment of the requirements of the degree of
Habilitation à Diriger des Recherches

Successfully defended: 04 December 2020

Jury :

President :	Serge FDIDA	-	Prof. Sorbonne University, FR
Reviewers :	Mérouane DEBBAH	-	Prof. CentraleSupélec & Huawei, FR
	James KUROSE	-	Prof. Uni. Mass.-Amherst, USA
	Emilio LEONARDI	-	Prof. Politecnico di Torino, IT
Examinators :	Françis BACCELLI	-	DR INRIA / ENS, FR
	Eduard JORSWIECK	-	Prof. TU Braunschweig, DE
	Luca MUSCARIELLO	-	Principal Engineer Cisco, FR
	Thrasyvoulos SPYROPOULOS	-	Prof. EURECOM, FR

Abstract (EN): This report presents the main findings of my scientific research after my PhD Thesis. During this period I have worked on three major topics, which span a wide spectrum of networking aspects: Cloud-RAN architecture for cellular networks, wireless edge-caching, and information diffusion in online social platforms. These relate to breakthroughs in communication networks during the last decade, considering the evolution towards 5G cellular networks and the proliferation of online social platforms. My research has resulted in original contributions in all three areas that advance the state of the art. In C-RAN architectures novel base station clustering mechanisms are proposed and the performance benefits from collaborative transmission is quantified for very large networks. In edge-caching new cache management policies are introduced that profit from multi-coverage, while user association and mobility is taken into account. Finally, in social networks an original mathematical model combines user posting activity with graph structure to exactly describe user influence inside online platforms. This research was made possible with the collaboration of several PhD and Master students.

Abstract (FR): Ce rapport présente les principaux résultats de mes recherches scientifiques après ma thèse doctorale. J'ai travaillé sur trois sujets principaux: l'architecture Cloud-RAN pour les réseaux cellulaires, les réseaux mobiles équipées avec des mémoires-cache, et la diffusion d'information sur les plateformes sociales en ligne. Ces domaines sont liés à l'évolution vers les réseaux cellulaires 5G et à la domination des plateformes sociales en ligne. Mes contributions originales dans les trois domaines ont fait progresser l'état de l'art. Pour les architectures C-RAN, je propose de nouveaux mécanismes de clustering et je quantifie les bénéfices de transmission collaborative dans les très grands réseaux. Pour les réseaux mobiles équipées avec des mémoires-cache, de nouvelles politiques de gestion de memoire sont introduites, qui profitent de la multi-couverture cellulaire; l'association des utilisateurs et la mobilité sont prises en compte. Enfin, dans les plateformes sociales, un modèle mathématique original combine l'activité des utilisateurs avec la structure du graphe sociale pour décrire précisément l'influence des utilisateurs sur les autres. Ma recherche a été rendue possible grâce à la collaboration avec plusieurs doctorants et stagiaires en master.

Keywords: wireless access, cooperation, Cloud-RAN, stochastic geometry, K-means, discrete optimisation, clustering, cellular networks, caching, wireless edge, Content Centric Networks, LRU, recommendations, Benders decomposition, mobility, online social networks, information diffusion, Markovian model, Newsfeed, Wall, balance equations, Twitter, Weibo, data, Markov Decision Processes

Contents

Chapter 1

Intro

In this report, I present the main results of my scientific research spanning the years I have worked as a CNRS Chargé de Recherche (CRCN), first affiliated with the Télécom ParisTech - LTCI lab (2013-2016) and later with the Sorbonne University - LIP6 lab (2017-now). This report supports my application to obtain the "Habilitation à diriger des Recherches" (HdR) title from Sorbonne University.

During this period, i worked on the general area of modelling, performance evaluation and optimisation of modern information networks, covering themes related to *wireless access, content delivery and information diffusion*. My main research is divided into three major topics, and I include in this report one chapter per topic:

- *Wireless access over Cloud-RAN*
  (Ch. 2, which covers the content from 1 journal, 5 conferences, 1 submission).
- *Caching at the wireless edge*
  (Ch. 3, covering the content from 3 journals, 5 conferences and 1 submission).
- *Post diffusion in social platforms*
  (Ch. 4, which covers the content of 1 conference and 1 submission).

The first two topics are related to the performance evaluation and optimisation of 5G cellular networks. They analyse specifically two distinct modern network features.

• (Ch. 2) The first feature allows cooperation between base stations; such cooperation is enabled by the novel cellular architecture of Cloud-RAN (C-RAN), which suggests that stations can be grouped together to form geographic clusters. The aim is to save energy consumption and to achieve coverage improvements by coordination of several base stations and the sharing of their available resources. The challenges addressed in my work are to find optimal clusters and to evaluate performance improvements through various ways of cooperative transmission.

• (Ch. 3) The second feature studied is the possibility to locally store multimedia content on wireless infrastructure (i.e. edge-caching), be that on base stations or on mobile devices. The aim is to avoid serving redundant requests through the backbone, and also to relieve the traffic charge of central data-centres. Hence, the edge becomes part of a global content-centric network architecture. This feature raises interesting questions related to optimal cache-management, user association, and storage memory leasing that are thoroughly dealt with in my research. The challenge in wireless edge-caching comes from the fact that receivers can be covered

by several wireless stations, hence it is important to determine what content to cache where, and how to route users to stations. A novel extension that I study is the improvement in caching networks through smart content recommendations.

More recently, I investigate social network platforms.

• (Ch. 4). The aim here is to introduce analytical models for post propagation that can sufficiently explain the underlying social platform mechanisms and to propose new ways of identifying high influencers. This research would further like to find ways to secure such networks by limiting the diffusion of fake news. The novelty here is that I introduce models which are dynamic over time and can combine the social graph structure with user posting activity over time. Such approach – inspired by queuing network analysis – is very original compared to the existing literature.

To obtain the research results presented in this report, I have collaborated with several PhD students: For (Ch. 2) with Luis David Alvarez-Corralles (PhD, 2014-2017 Télécom ParisTech). For (Ch. 3) with Jonatan Krolikowski (PhD, 2015-2018 Télécom ParisTech), Chedia Jarray (PhD, 2016-2019 ENIG Gabes, Tunisia), and Theodoros Giannakas (PhD, 2017-2019 EURECOM). For (Ch. 4) a new PhD student is recruited, Ricardo Jose Lopez Dawn (2020-..., Sorbonne University - LIP6).

My research has also been supported by several Master students, with whom we have published in very important conference venues. For (Ch. 2) Zakarya Boubazine (LIP6), Bahiya Chakiri (LIP6), Liticia Touzari (LIP6) and Hanane Djeddal (LIP6). For (Ch. 3) Apostolos Avranas (Télécom ParisTech). For (Ch. 4) Antoine Vendeville (LIP6).

I have obtained by myself all the financial resources necessary for my research: All Master stages have been covered by local projects from Télécom ParisTech and the Sorbonne-LIP6. For (Ch. 2) the PhD thesis of L.D. Alvarez-Corrales was funded by a Futur&Ruptures Mines-Telecom project. For (Ch. 3) the thesis of J. Krolikowski was supported by a Digiteo-Digicosme Paris-Saclay project. For (Ch. 4) I have obtained an ANR Jeune Chercheur JCJC project (2020-2024) that finances fully two PhD theses.

Chapter 2

Wireless access over Cloud-RAN

Contents

The first part of my work (2012-today) considers 4G and 5G cellular architectures, where the base stations can be grouped together to form collaborating clusters of wireless transmission. This research is gathered under the umbrella of C-RAN architectures and considers clustering and cooperative transmission problems. I supervised 1 PhD student and 5 Masters theses related to the content of this chapter.

2.1 Challenges of Cloud-RAN

Mobile network data traffic is growing explosively due to multimedia consumption from smartphones and tablets, which are constantly connected to the Internet. The end-user needs are continuously growing as shown in the Cisco report growth of data traffic (2018-2023) [Cisco 2020]. The report predicts a 10% compound annual growth rate of number of devices and connections, with an emphasis on Machine-to-Machine Internet-of-Things (M2M-IoT). The average smartphone connection speed should increase by a 4-fold until 2023 to satisfy the demand.

Mobile network operators need to improve their coverage, throughput, as well as data processing capacity, and this has been a major focal point during the 4G and 5G research and development period. The major concern has been for cell-edge users and more generally areas with high interference. Several efforts have

been directed towards densification [Zhang *et al.* 2017], i.e. the deployment of more base stations and small cells for a given geographic area, forming a so-called heterogeneous network (HetNet) [Andrews *et al.* 2014]. This solution aims to bring the user equipment closer to the cellular access point, but it has high deployment cost (CAPEX and OPEX) as the network scale grows. One needs to take also into account the energy consumption of the densified mobile network infrastructure, which is already very high and grows with the number of installed stations.

From a signal processing and transmission perspective, novel technics are applied to increase the spectral efficiency such as multiuser MIMO, and massive MIMO [Jungnickel *et al.* 2014]. All these techniques can increase the capacity of an isolated cell, but often come at a cost of high inter-cell interference.

To optimise deployment cost and energy consumption in ultra dense networks, and to furthermore apply technics of interference mitigation efficiently, an evolution of the mobile network architecture has been envisioned and implemented in the recent years: the *Cloud Radio Access Network (Cloud-RAN or C-RAN)*. The idea comes naturally from the way a base station is structured. A typical base station is split into a radio unit and a signal processing unit. The Remote Radio Head (RRH) provides the interface to the fiber and performs digital/analog conversion. The baseband signal processing is performed at the BBU. The RRH and BBU are connected via optical fiber or microwave connection, and their distance can be extended up to several kilometres, where the limitation is due to processing and propagation delay [Checko *et al.* 2015]. This two-part separation of a base station gave rise to the idea [Lin *et al.* 2010] that the BBUs of several stations can be grouped together and be gathered in a single location called BBU pool; this way they can be shared among several low-cost, low-complexity RRHs, which are geographically scattered [China-Mobile 2011]. This locally centralised architecture is given the name C-RAN, because the BBU resources form a processing cloud that serves several stations. This cloud now serves the total coverage area of the RRH group.

The locally centralised C-RAN architecture allows network information from several sites to be shared at the BBU pool, a novelty which enables cooperative communication techniques such as network coordination, joint resource management, and coordinated multipoint (CoMP) [Wu *et al.* 2015], [Pan *et al.* 2018], which are not possible unless low delays are guaranteed for message exchange between BBUs. These techniques can efficiently mitigate the co-channel interference within each C-RAN cluster. Furthermore, the computing resources of the BBU pool can be controlled as an ensemble and adapt to non-uniform fluctuations of traffic between base stations, which may be heavily or lightly loaded during different hours of the day [Chen *et al.* 2018]. This results in utilising fewer BBUs compared to the traditional architecture, and consequently in decreasing the cost of network operation (energy efficiency). Another important benefit of the C-RAN architecture is that intra-BBU pool handover delays are reduced and network performance is increased.

My research has focused on two key sub-problems related to C-RAN. These two problems are not independent the one from the other.

1. **Base-station clustering:** how to appropriately group base stations of an area into clusters, each cluster having its own shared BBU-pool.

2. **Cooperative transmission:** how to achieve higher network performance (coverage, throughput) by appropriate choice of the transmission scheme, given the clustering.

2.1.1 Clustering

In general, there are two ways to determine groups of stations. In my research, presented in this chapter, I have studied problems related to both types of clustering to understand their potential benefits and also their performance differences.

2.1.1.1 Disjoint clustering

In the disjoint clustering, the RRHs in the region are partitioned into several non-overlapping clusters, e.g. as in [Akoum & Heath 2013]. Then we are looking for clusters $\mathcal{C}_1, \ldots, \mathcal{C}_K$ with the property that $\mathcal{C}_i \cap \mathcal{C}_j = \emptyset$ and $\bigcup_{k=1}^{K} \mathcal{C}_k = \{$ the set of all stations $\}$. Sub-sets of stations within a created cluster can apply various cooperative transmission techniques – including CoMP – to serve the users that move inside the cluster coverage area. The cluster edge users can still suffer from inter-cluster interference, assuming that no information is exchanged between clusters.

2.1.1.2 User-centric clustering

In the user-centric cluster, each user is individually served by its own nearby RRHs [Nigam *et al.* 2014], [Tanbourgi *et al.* 2014], [Pan *et al.* 2018]. Now, the user is the center of its own cluster. Hence, there is soft partition of the stations; clusters overlap with each other, serving same sub-sets of users. This approach does eliminate the potential edge-effects of hard clustering, but it requires increased message exchange and cooperation between all possible pairs of stations. Hence, the economy of resources envisioned by the Cloud-RAN architecture may not be guaranteed here, although communications performance can significantly be improved.

• **Features:** In both cases, the clusters can be determined by various network features. The basic feature is the Euclidean distance between stations. But, since the main problem is a wireless one, other features can be included related to traffic demand, availability of resources, and user Quality-of-Service (QoS) satisfaction. My work has tested all these variations in different models.

2.1.2 Cooperative transmission

A major issue in full-frequency reuse traditional cellular networks has been the low coverage and low service quality at the cell edge, due to inter-cell interference. To deal with this, different stations can coordinate their transmission, or

even exploit interference in a constructive way through coherent cooperation. Cooperative transmission can be achieved at the cost of tighter synchronisation and increased overhead, and for this the BBU-pool architecture is most appropriate [Karakayali *et al.* 2006], [Gesbert *et al.* 2007]. Note here that, users close to the cell-centre do not suffer from interference, and there is no need for coordination. Hence, coordination can be selectively applied on different geographic areas of the same cluster, and not everywhere.

In my research I have focused mostly on cooperation in the cellular downlink, which can improve average as well as cell-edge throughput. In the downlink, there are the following standard cases of cooperative transmission [Irmer *et al.* 2011].

2.1.2.1 Coordinated scheduling/beamforming

The user is connected and served from just one serving RRH, but user scheduling and beamforming decisions are coordinated among the RRHs with the aim to minimise interference. This is usually called Inter-cell Interference Coordination (ICIC). The simplest implementation for this scheme is when neighbouring RRHs are silenced on certain frequency bands in order to reduce interference for the targeted user under service by a single RRH.

2.1.2.2 Joint-processing CoMP

Data for the same user is available and transmitted over several RRHs of the network in order to exploit interference paths constructively. A simple extension of the above ICIC scheme, is when the serving RRH is not fixed but rather dynamically picked from a potential set of transmitters, based on channel quality criteria (dynamic cell selection). The most advanced scheme is when full channel state information (CSI) and user data is available to all stations, which jointly and coherently transmit to one user, thus performing Joint Transmission (JT). This increases the Signal-to-Interference-plus-Noise Ratio (SINR) at the mobile and offers higher achievable bit rates. In general, performance of different transmission policies varies depending on the amount of CSI and data exchanged among stations inside the cluster.

Typically, precoding is applied in order to combine transmit signals from all cooperative RRHs that serve a user.

2.2 Modelling ingredients

2.2.1 Signalling

The first ingredient for the analysis is to give expressions for the useful signal as well as the interference received by a user (he/his) who is served by a set of RRHs.

Specifically, let a typical user be placed at the centre of coordinates $(0, 0)$. Then, suppose for simplicity that $M = 2$ single antenna RRHs are placed at a distance r_1 and r_2 from the user, with $r_1 < r_2$. Let $p > 0$ [Watt] be the signal transmission power per RRH. Each transmitted signal experiences random fading $h_1 > 0$, $h_2 > 0$

(with mean 1) and path-loss with exponent $\beta > 2$. Also, suppose that the user's receiver experiences noise with power σ^2 [Watt].

- In the standard cellular case *without cooperation* [Andrews *et al.* 2011], the user is served by just one station (cell) and the signal from the second one is experienced as noise,

$$SINR_{NO}(r_1, r_2) = \frac{ph_1 r_1^{-\beta}}{ph_2 r_2^{-\beta} + \sigma^2} \quad \text{[No coop]}. \tag{2.1}$$

  During this time, the second station is serving a user in its cell using power p.

- When the two RRHs *coordinate their scheduling*, the second station can be silenced to reduce interference for the user in cell 1,

$$SINR_{CO}(r_1, r_2) = \frac{ph_1 r_1^{-\beta}}{\sigma^2} \quad \text{[Coordinate]}. \tag{2.2}$$

  During this time, the second RRH does not serve users over the silenced sub-band. Interference is reduced locally at the cost of global network service. This scheme has reduced energy consumption.

- When the two RRHs perform *joint transmission*, information over h_1, h_2 and user data is available to both RRHs. The signal of each station is then weighted by appropriate pre-coding weight using this channel information, to achieve coherent addition at the receiver and result in

$$SINR_{JT}(r_1, r_2) = \frac{ph_1 r_1^{-\beta} + ph_2 r_2^{-\beta}}{\sigma^2} \quad \text{[JT]}. \tag{2.3}$$

  During this time, the second RRH cannot serve other users over this sub-band, because it collaborates with the first station for the service of a single user. Energy consumption is the same as [No coop].

  Note here, that more generally, the pre-coded joint transmission by the two stations can be written as

$$SINR_{JTw}(r_1, r_2) = \frac{\left| w_1 \sqrt{ph_1 r_1^{-\beta}} e^{i\theta_1} + w_2 \sqrt{ph_2 r_2^{-\beta}} e^{i\theta_2} \right|}{\sigma^2} \quad \text{[JTw]}, \tag{2.4}$$

  where w_1, w_2 are complex valued weights that depend on the channel gains and channel angles $(h_1, \theta_1, h_2, \theta_2)$. Various choices of the weights lead either to (2.3) or to some other expression for the cooperative signal reception.

2.2.2 User association

In standard cellular networks, a user is associated with one RRH station, either the geographically closest one or the one with the strongest received signal. In the C-RAN context with CoMP, users can be assigned for joint transmission to a group of

stations. The association rule to a group of RRHs creates complicated scenarios of interference, as it has many degrees of freedom. Also, the association will strongly depend on the clustering scheme:

- if the clustering is *user-centric*, the user can choose the *closest* stations to collaborate for his service. In this case the most important question is how many stations should collaborate for optimal global performance.
- if the clustering is *disjoint*, the user can only choose among the stations of his own cluster. Here, there might be cases where the RRH options are not the most favorable ones for the user, e.g. when the user lies at the cluster-edge. In these cases the inter-cluster interference is strong. An important question, aside the size of the clusters and the clustering method, is to evaluate the performance degradation compared to the user-centric clustering.

Another important issue related to user association is load balancing. To achieve an equally fair consumption of resources among clusters and to balance inter-cluster interference, load balancing algorithms can determine the association rule among clusters.

2.2.3 Node positions

In the expressions for signalling, the received SINR is a function of the distances r_1 and r_2 from the serving (and/or interfering) RRH to the user. These distances depend on how the relative position of users and stations on the plane are modelled. The choice to model positions is critical for the solution of our problem and the derived performance. In my research I have worked with two types of positioning.

- **Deterministic positioning:** In this case the positions of RRHs and users on the plane are pre-determined and fixed. The performance of the network is derived each time for a given set of positions.
- **Random positioning:** The positions of RRHs and users exhibit spatial randomness. In this case, the performance of the network can be derived "on average" over all possible positions and their distribution. This is interesting because the solution is insensitive to a specific given constellation. But it becomes dependent only on the position statistics. Most often, the Poisson point process (PPP) distribution is used to model 2D-randomness.

The use of Poisson point processes has become a standard approach over time, because it can give closed formulas for average network performance, something that in the past was possible only through large-scale network simulations (e.g. NS3). The closed formulas provide a more concrete interpretation of the interactions between model parameters. The standard paper that introduce PPPs in the cellular literature is [Andrews *et al.* 2011] and the tutorial book is [Baccelli & Blaszczyszyn 2010]. The simplicity of the PPP lies in the fact that the number of nodes within a finite

2D window is Poisson distributed with mean value proportional to the area surface. Given the number of nodes, these are distributed in the area uniformly at random over the two dimensions.

2.3 Optimal disjoint clusters with QoS constraints

This subsection presents the following research work (published in 2012), which was my first publication on the topic. It was done in collaboration with Qualcomm, Germany, during my post-doc at Zuse Institute Berlin (ZIB), Germany.

Publication [Giovanidis *et al.* 2012]: A. Giovanidis, J. Krolikowski and S. Brueck, "A 0-1 program to form minimum cost clusters in the downlink of cooperating base stations", *IEEE Wireless Communications and Networking Conference (WCNC)*, 2012

In this paper, I aimed to optimise jointly the clustering, and user association problem, while considering signalling with JT-CoMP. The work considers deterministic user and base station positions that are known a-priori. The starting point is a cellular placement of M single antenna stations indexed by m, and N users indexed by n. The question is "how should the stations be grouped into clusters and how should the users be associated to them", while satisfying some SINR quality constraint. The problem allows for CoMP joint transmission, when a user is associated to a cluster. The solution finds a disjoint clustering scheme. Similar problems were studied later, in [Sanguinetti *et al.* 2016] and [Li *et al.* 2015] that included also optimisation over the pre-coders. In my work, the pre-coders were assumed binary (serve or not). Two types of binary random variables are introduced:

- The *user association* variables between each user – station pair $\{x_{n,m}\}$. When $x_{n,m} = 1$ the user n is served by station m, else he is not.
- The *station cooperation* variables $\{y_{m,\ell}\}$ for all pairs of stations $m < \ell \leq M$ ($y_{m,\ell} = y_{\ell,m}$). When $y_{m,\ell} = 1$ the two stations are in the same cluster.

To achieve disjoint clustering, we introduce a constraint over the unknowns that enforces all collaborating stations to form *complete subgraphs*, i.e. that all $y_{m,\ell}$ of nodes belonging to a cluster are 1, and all $y_{m,k}$ of nodes in different clusters are 0.

The constraint reads,

$$y_{m,\ell} + 1 \geq y_{m,k} + y_{k,\ell} \qquad \text{for all } (\mathrm{m}, \mathrm{k}, \ell) \text{ triplettes.} \tag{2.5}$$

To understand this, take $M = 3$, then we get the set of inequalities

$$\begin{aligned} y_{1,2} + 1 &\geq y_{1,3} + y_{3,2} \\ y_{1,3} + 1 &\geq y_{1,2} + y_{2,3} \\ y_{2,3} + 1 &\geq y_{2,1} + y_{1,3} \end{aligned}$$

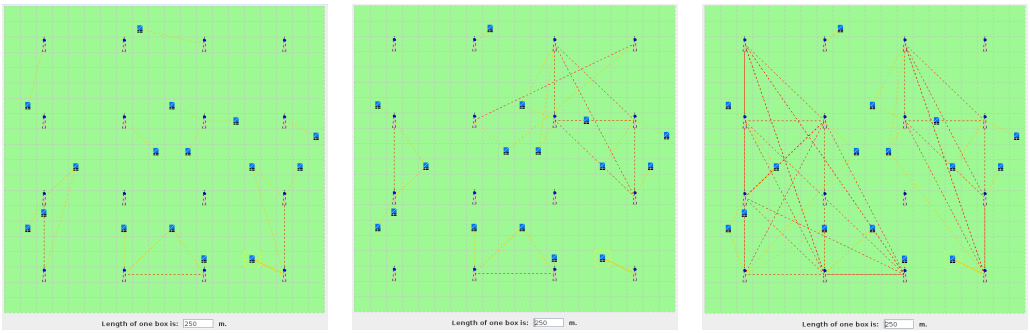

Figure 2.1: Increase in the QoS requirements spreads cooperation.

If we assume that $y_{1,2} = 1$, i.e. that the two stations (No."1" and No."2") are in the same cluster, then the set of inequalities gives necessarily $y_{1,3} = y_{2,3}$. Then both variables are either 0 or 1. So, either there are two disjoint complete clusters $\{\{1,2\},\{3\}\}$, or one complete cluster $\{\{1,2,3\}\}$. It is not possible that two edges are active but not the third. This property is kept for all clusters with larger size. Furthermore, the larger the cluster, the larger the $\sum_{m,\ell\in\mathcal{C}} y_{m,\ell}$.

A second set of inequalities, guarantees that for the service of a single user n, only stations from the same cluster can collaborate. We require,

$$y_{m,\ell} + 1 \geq x_{n,m} + x_{n,\ell} \qquad \text{for all users n, RRH pairs } (\mathrm{m},\ell),\ m \neq \ell. \tag{2.6}$$

If we set above $y_{m,\ell} = 0$, i.e. that stations m and ℓ belong to disjoint clusters, then $x_{n,m} \neq x_{n,\ell}$, i.e. the user can be associated either to one station or the other but not both. If $x_{n,m} + x_{n,\ell} = 2$, then necessarily the two stations cooperate in the same cluster $y_{m,\ell} = 1$.

With the above two inequality sets, both the clustering as well as the user association to clusters are disjoint sets. We further require that the user association satisfies the Quality-of-Service (QoS) constraints for each user $n = 1,\ldots,N$. This is defined by their SINR, which is a function of the assignment variables x,

$$SINR_n = \frac{\sum_{m=1}^{M} x_{n,m} \cdot ph_{n,m} r_{n,m}^{-\beta}}{\sum_{m=1}^{M} \left(\sum_{v=1, v\neq n}^{N} x_{v,m} \cdot ph_{n,m} r_{n,m}^{-\beta} \right) + \sigma_n^2} \quad \geq \quad \Gamma_n. \tag{2.7}$$

The SINR formulation above unifies and generalises the expressions in (2.1)-(2.3) to include $M > 2$ stations and $N > 1$ users, and allow for various signalling and user association rules. Signals from non-serving stations will be treated as interference. Each station has a different distance $r_{n,m}$ and fading $h_{n,m}$ value per user n. The above formula allows for service of a single user by several stations in clusters, transmitting with CoMP.

We further allow for a station to transmit to several users on the same frequency band, by superposition coding. To limit the number of users associated per station,

we include our final constraint with L integer

$$\sum_{n=1}^{N} x_{n,m} \leq L, \qquad \text{for all stations m} = 1, \ldots, \text{M}. \tag{2.8}$$

The aim of the work at the time was to minimise the *modification cost* of the traditional cellular network subject to QoS constraints. The cost was defined as the sum over all collaborative pairs, because my idea at the time was that an optical fibre should connect all pairs of nodes together to exchange the necessary information to serve as a cluster. This defines the following discrete binary optimisation problem

$$\begin{array}{ll} \min_{\{x\},\{y\}} & \sum_{m=1}^{M} \sum_{\ell=1,\ m\neq\ell}^{M} c_{m,\ell} \\ \text{subject to} & (2.5),(2.6),(2.7),(2.8) \\ & x_{n,m} \in \{0,1\},\ y_{n,m} \in \{0,1\}, \text{ for all } n, m \end{array} \tag{2.9}$$

The program guarantees that all feasible clusters are disjoint and that a feasible solution assigns a user to exactly one cluster. Also the solution has a $min - max$ behaviour for the optimal cluster size. A simulator was developed to randomly place users and stations on a plane. The open source mixed integer problem solver SCIP, which implements Branch-and-Bound, was used to derive the $0-1$ solution to the Prob. 2.9. An example of the clustering and user association solution is shown in Figure 2.1 for growing $SINR$ threshold values of $\Gamma_n = \Gamma$ (left to right).

2.4 Performance of user-centric pair clusters

The second topic introduces randomness in user and station positions. Both are modelled as Poisson point processes (PPP). This work was done in collaboration with F. Baccelli during my post-doc at INRIA, France. Following the groundbreaking paper [Andrews *et al.* 2011], which introduced PPPs as a standard tool for performance analysis in cellular networks, our work derived first results on stochastic geometry modelling of cooperation in wireless networks. Two related publications:

Publication [Baccelli & Giovanidis 2013]: F. Baccelli and A. Giovanidis, "Coverage by pairwise base station cooperation under adaptive geometric policies", *Asilomar Conference on Signals, Systems and Computers*, 2013.

Publication [Baccelli & Giovanidis 2015]: F. Baccelli and A. Giovanidis, "A Stochastic Geometry Framework for Analyzing Pairwise-Cooperative Cellular Networks", *IEEE Transactions on Wireless Communications*, 2015.

The aim of this research is to quantify the coverage improvements of user-centric clustering. At the time (2012-2013) there was confusion about the extent of improvement by CoMP in cellular networks. We conclude in this work that on average the SINR at a random user position increases by 17% (for the specific

signalling scheme studied) through downlink cooperation between the two closest stations of the user. Our work influenced other researchers who extended our results [Tanbourgi *et al.* 2014] and [Nigam *et al.* 2014].

To answer this question, we considered a random network with PPP distribution of stations having density λ [stations/m^2] and focus on the typical user o at the Cartesian origin. This is the "average" user, and is representative of any other in the network. The starting point is the standard cellular architecture, where each station is responsible for the service of one user at its cell. The power budget globally spent for the service of each user is fixed at $p > 0$ [Watt].

We allow either full service of the user by its geographically closest station, or service in cooperation: In the second case, the user chooses *the first and second closest station* to cooperate for his service. The total power budget is split, and $p/2$ [Watt] is sent by the first closest station, whereas $p/2$ [Watt] by the second closest station (we wanted to compare the schemes in a fair way related to power consumption). The signals are pre-coded in a specific way to add constructively at the user's receiver in a scheme known as Willems' encoding. The SINR at the typical user o takes the following expression,

$$\begin{aligned} SINR_o(r_1, r_2; \rho) &= \frac{p h_1 r_1^{-\beta}}{\sigma^2 + \mathcal{I}(r_2, \rho)} \mathbf{1}_{\{r_1 \leq \rho r_2\}} \\ &+ \frac{\frac{p}{2} h_1 r_1^{-\beta} + \frac{p}{2} h_2 r_2^{-\beta} + p\sqrt{h_1 h_2 r_1^{-\beta} r_2^{-\beta}}}{\sigma^2 + \mathcal{I}(r_2, \rho)} \mathbf{1}_{\{r_1 > \rho r_2 \ \& \ r_1 \leq r_2\}} \end{aligned} \quad (2.10)$$

To clarify the above expression, we have introduced a position switch to ask for cooperation between stations, or not: When the user is close to the center of the cell (geographically closest) his position satisfies $r_1 \leq \rho r_2$, and there is no need for cooperation because first-order interference is low. When the user is close to the border of the cell his position satisfies $r_1 > \rho r_2$ & $r_1 \leq r_2$, then interference is high, and CoMP can treat the strongest interference from the second closest station as beneficial signal. The parameter $\rho \in [0, 1]$, which can be tuned, determines how close to the cell-edge the user should be to ask for cooperation, see Figure 2.2. Willems' cooperation is a CoMP scheme that enables coherent addition of the signals at the receiver side. It results from (2.4) by setting $w_1 = \frac{e^{-i\theta_1}}{\sqrt{2}}$ and $w_2 = \frac{e^{-i\theta_2}}{\sqrt{2}}$, hence only channel phase information is needed (which is less than full-JT). The beneficial term is $p\sqrt{h_1 h_2 r_1^{-\beta} r_2^{-\beta}}$, and it should overcome the loss $\frac{p}{2}\left(h_1 r_1^{-\beta} - h_2 r_2^{-\beta}\right)$ due to service with $p/2$ [Watt] from a station further away than the closest one.

In the SINR expression the interference part is treated as noise. The interference is the sum of all signals destined towards all users other than the user of interest (typical user). Then, for each other user in the network its two closest stations will cooperate or not depending on his relative position to these stations, following the policy with global parameter ρ. The coherence term for cooperating signals is 0 in the interference, because coherence is scheduled by every cluster for its own served user, not the one at the Cartesian origin. Since we work with PPP we get an

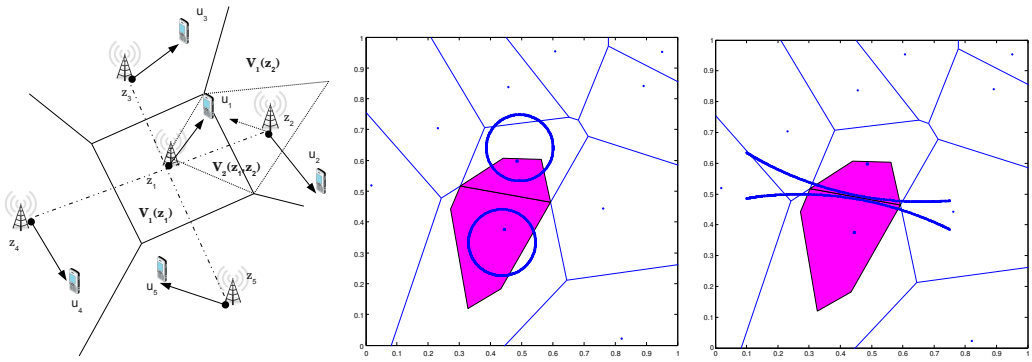

Figure 2.2: Cooperation of two closest stations and regions for $\rho = 0.4$ and $\rho = 0.9$. The shaded area is the 2-Voronoi cell, i.e. the geometric locus of all points having the specific two stations as 1st and 2nd neighbour. The users asking for cooperation lie outside the circles.

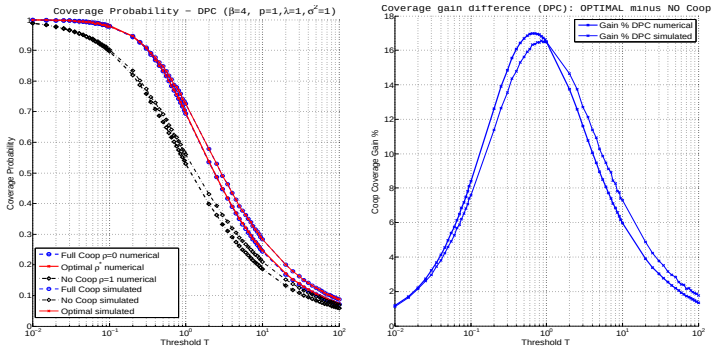

Figure 2.3: SINR coverage benefits from user-defined cooperation in pairs (CCDF).

interference sum with infinite summands, one per network user,

$$\mathcal{I}(r_2, \rho) = \sum_{u \neq o} \left[p h_{u,1} r_{u,1}^{-\beta} \mathbf{1}_{\{\text{no coop u}\}} + \frac{p}{2} \left(h_{u,1} r_{u,1}^{-\beta} + h_{u,2} r_{u,2}^{-\beta} \right) \mathbf{1}_{\{\text{coop u}\}} \right], \quad (2.11)$$

where $(r_{u,1}, r_{u,2})$ is the pair of distances from the 1st and 2nd closest neighbour of user u to the typical user o at the cartesian origin. We note $\mathcal{I}(r_2, \rho)$ because the interference to the typical user comes from the second closest station at r_2 and from all other stations further away. We assume that intra-cell interference can be nulled-out by dirty paper coding in the serving cell at r_1.

The SINR expression in (2.10) is a function of the pair of distances (r_1, r_2). Given that the positions of stations follow a 2D Poisson point process, the joint distribution of the random distances (R_1, R_2) from the origin can be derived to be

$$f_{R_1,R_2}(r_1, r_2) \quad = \quad (2\pi\lambda)^2 r_1 r_2 e^{-\lambda\pi r_2^2}. \quad (2.12)$$

The expected value of distance R_1 is $\mathbb{E}[R_1] = \frac{1}{2\sqrt{\lambda}}$, and the expected value of distance R_2 is $\mathbb{E}[R_2] = \frac{3}{4\sqrt{\lambda}}$. Another useful expression is that the probability for a user not

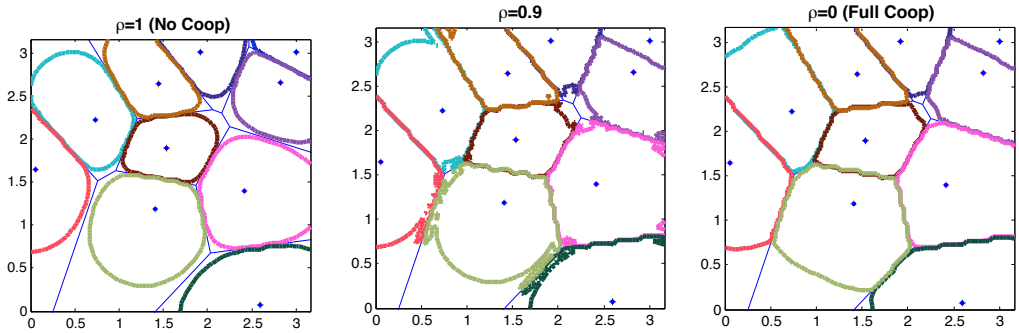

Figure 2.4: Increase in coverage regions due to cooperation, with ρ.

to demand for cooperation is the square of the ρ parameter

$$\mathbb{P}[\text{No Coop}] = \mathbb{P}[r_1 \leq \rho r_2] \quad = \quad \rho^2. \tag{2.13}$$

Since the joint probability distribution is available in (2.12) we can now calculate the CCDF (complementary cumulative distribution function) of the SINR for the "average positioned user" (i.e. *typical user*) and characterise completely the offered QoS of the cooperation scheme

$$\mathbb{P}(SINR(\rho) > T) \quad = \quad \mathbb{E}\left[\mathbb{P}(SINR(r_1, r_2; \rho) > T) \mid R_1 = r_1, R_2 = r_2\right]. \tag{2.14}$$

The above expression, can be simplified to an integral over the product of the Laplace transform from the received signal and the Laplace transform from the interference. Figure 2.3 plots the CCDF for the SINR and quantifies the coverage benefits of the scheme (17% max) compared to the scheme without cooperation. We also show how the coverage area increases as a function of ρ, in Figure 2.4.

2.5 Performance of disjoint pair clusters

The third topic covers a large part of the doctoral thesis under my supervision:

• **Ph.D. thesis** Luis David Álvarez-Corrales [Álvarez-Corrales 2017]): "Cooperative communications in very large cellular networks", Co-supervised with Philippe Martins. Télécom ParisTech, EDITE 2017.
Project financed by Fondation Télécom - Futur & Ruptures.

The thesis was realised during 2014-2017 while I was with Télécom ParisTech. The main idea was to introduce disjoint clusters (instead of user-defined) in random geometries of node positions, and evaluate the performance benefits from cooperative transmission. It is somehow a natural extension of my previous topic, but the restriction on disjoint clusters eventually made the analysis much more difficult for reasons to be explained. Before launching the thesis I had performed a literature survey with first investigations for disjoint clustering in random geometries. It can

be found archived as unpublished Technical report in [Giovanidis 2016]. In this chapter i will cover the results from two published conference articles. The articles show an average performance gain of 14% in cellular coverage for cooperation from disjoint pairs in the downlink.

Publication [Giovanidis *et al.* 2015]: A. Giovanidis, L. D. Álvarez-Corrales and L. Decreusefond, "Analyzing interference from static cellular cooperation using the Nearest Neighbour Model", *13th Int. Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOpt)*, 2015.

Publication [Álvarez-Corrales *et al.* 2016]: L. D. Álvarez-Corrales, A. Giovanidis and P. Martins, "Coverage Gains from the Static Cooperation of Mutually Nearest Neighbours", *IEEE Global Communications Conference (GLOBECOM)*, 2016.

We model again the station locations by a stationary PPP Φ with density λ [stations/m^2]. A user terminal is assumed at the cartesian origin $(0,0)$ and we examine the network performance there. This is again the "typical user" approach. The stations are organised in static clusters, i.e. their groups are pre-defined irrespective of the user positions and stay unchanged. The method to determine groups is based on node proximity: stations in a small distance from each other will tend to create interference to one another, hence they have interest to collaborate. We apply here the Nearest Neighbour model for clustering [Häggström & Meester 1996], which states that two stations belong to the same group if one of the two is the Nearest Neighbour (NN) of the other. Given a station at position x, its NN is the station at position y with minimum distance from x, i.e. such that

$$y \quad = \quad \arg\min_{z\in\Phi\setminus\{x\}} d(x,z), \tag{2.15}$$

where $d(x,z)$ is simply the 2D Euclidean distance. We denote this relation with the arrow $x \to y$. We show an instance of such clustering in Fig. 2.5 (left). Observe that all clusters are of finite size; since clusters are formed by the relative distance between their nodes, the size of a cluster can vary depending purely on geometry.

To keep the analysis simple we restrict ourselves to a variation of the above model, where only pairs of nodes or single isolated stations can exist. These are defined in the following way

- **Cooperating pair (x,y)**: The two stations cooperate if the one is the NN of the other, i.e. $x \to y$ and $y \to x$. We then denote the pair by $x \leftrightarrow y$.
- **Single station (w)**: A station w does not cooperate, when its own NN has some other station as NN, and we denote this by $w_{\#}$.

We explain geometrically the two definitions in Fig. 2.5 (middle) and show a specific realisation of such a process of singles and pairs with their surface of association (Voronoi cell) in Fig. 2.5 (right).

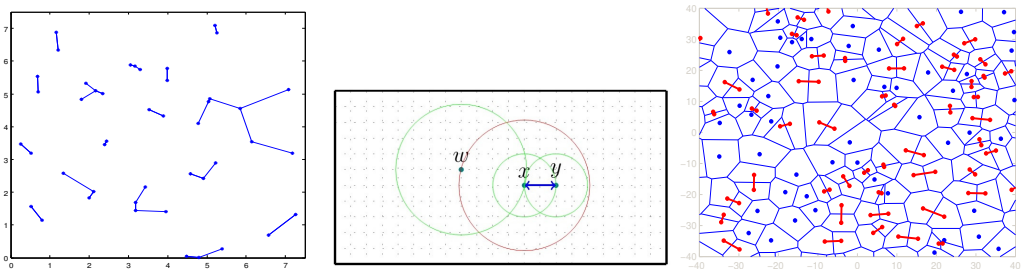

Figure 2.5: The Nearest Neighbour graph: general case and singles-pairs.

The way we have defined the singles and pairs, can determine two new spatial point processes, one for the singles (by removing all pairs) Φ_s and one for the pairs (by removing all singles) Φ_p. We summarise our findings for the two new processes:

- The process of singles Φ_s and the process of pairs Φ_p are not Poisson, but are stationary.

- The density of pairs is $\lambda\delta$ and the density of singles is $\lambda(1-\delta)$ where δ is a constant that does not depend on the original density λ. In fact

$$\delta = \frac{1}{2-\gamma} \approx 0.62, \qquad \gamma = \frac{2}{3} - \frac{\sqrt{3}}{2\pi}, \tag{2.16}$$

  where $\gamma\pi$ is the surface of intersection of two discs with unit radius and centres lying on the circumference of each other. Hence, 62% of nodes are in pair, and 38% are singles, irrespective of λ.

- The distance between the nodes of one pair is a random variable and is Rayleigh distributed

$$P(d(x,y) \leq r \mid x \leftrightarrow y) \quad = \quad 1 - e^{-\lambda\pi r^2(2-\gamma)}. \tag{2.17}$$

Given the fact that the two resulting processes for the clusters are not Poisson, we have difficulties in working analytically. We have been able to derive expressions for the mean interference from all singles and from all pairs, as perceived by the typical user at $(0,0)$, but an exact $SINR$ analysis is difficult.

Approximation: We use an *approximative model*, based on the above observations to derive further performance results: We approximate the process of singles by a Poisson process with density $\lambda(1-\delta)$, following (2.16). We approximate the process of pairs by a Gauss-Poisson process. The parent nodes are distributed as a Poisson process with density $\frac{\lambda\delta}{2}$ and each parent has exactly one child node. The child node is distributed uniformly in angle around the parent node, and in a distance Rayleigh distributed as in (2.17). This way, the approximative model mimics the properties of the two processes Φ_s and Φ_p, while keeping the Poisson property to allow for calculations.

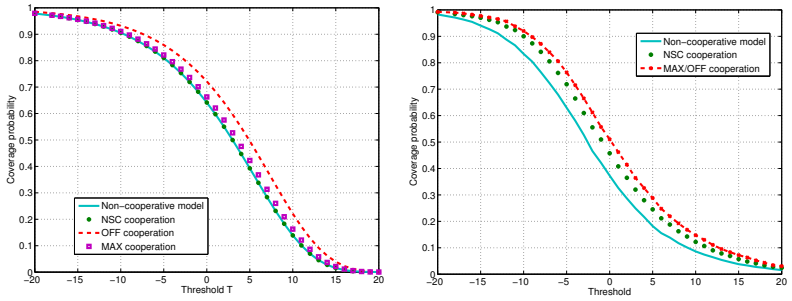

Figure 2.6: Coverage Gains from NN clusters of pairs & singles and various signalling schemes: association to a fixed station (left), association to the closest cluster (right).

In the approximative model the positions of the singles are Poisson distributed and the distribution of their distance from $(0,0)$ is known. The same goes for the statistics of the distances for the parents of pairs being also a Poisson process. It thus remains to find the distribution of the distance from the child per parent to the origin (0,0). Given that the parent is at distance r from the origin, the distance of its child from the typical user is found to be *Rice distributed* with parameters (r, α). Its probability density function is

$$f_Z(z \mid R = r) \quad = \quad \frac{z}{\alpha^2} e^{-\frac{z^2+r^2}{2\alpha^2}} I_0\left(\frac{zr}{\alpha^2}\right), \tag{2.18}$$

where $\alpha = (2\lambda\pi(2-\gamma))^{-1/2}$ and $I_0(x)$ is the modified Bessel function of the first kind with order zero. With this result, we can thus determine the joint distribution of the two distances from the closest pair-cluster to the origin $f_{R2,Z}(r,z)$, thus establishing a result similar to (2.12) we had in the user-centric case.

Having determined the joint distribution of distances from a pair, we can characterise the interference coming from single stations in Φ_s as well as from the pairs which might cooperate in various ways, as in (2.2), (2.3) or (2.4). We can again calculate the CCDF for the "typical user", using the Laplace transforms of the signal and the interference. We evaluate two cases. (Case I) We fix a single station at distance r_o to serve the user, and all other stations that might cooperate in pairs or not are interference. (Case II) The user is served by the geometrically closest cluster (single or pair) and all the rest are interference. The coverage probability is plotted in Fig. 2.6 for Case I (left) and Case II (right) where we observe consistently coverage benefits of up to 14%. The highest benefits are observed for the case OFF where one of the two stations in pair does not transmit, and generates minimum interference. The case NSC refers to the case where both stations from the pair serve the same user, and their signals are added coherently, as in (2.3).

2.6 Traffic-aware static clustering

In the random position models so far, we have considered the relative position of nodes as the most important feature to determine clusters of stations. This as-

sumption is motivated by two facts: the inter-cell interference is strongest between neighbouring stations, and also CoMP can be beneficial for regions that are covered by all stations in the cluster. Realistically, however, there is also another very important factor to determine the usefulness of cooperation: traffic, and consequently resource scarcity and availability. We would like to include, in addition to relative distance, a traffic-aware criterion, which should have the following impact:

- When two stations are complementary in resources, i.e. the one has scarce resources the other abundant (equiv. the one has high traffic demand and the other low traffic), then they should cooperate, so that the one helps the other.
- When both stations have abundant resources (equiv. low traffic demand), it is irrelevant whether they cooperate or not.
- When both stations have scarce resources (equiv. high traffic demand), they should belong to separate clusters, i.e. they should not collaborate.

The challenge here is to combine the Euclidean distance in $\mathbb{R}^2$ between stations at (x_1, y_1) and (x_2, y_2),

$$d_{E2}(x_1, y_1, x_2, y_2) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}, \tag{2.19}$$

with a relative difference between the third feature that counts the available resources z_1 and z_2. Note that the more the traffic that asks service from a station, the less resources are available at this station.

The first trivial idea is to consider a triplette of features per station $u = (x_1, y_1, z_1)$ and $v = (x_2, y_2, z_2)$ and generalise the Euclidean distance in 3D,

$$d_{E3}(x_1, y_1, z_1, x_2, y_2, z_3) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2 + (z_1 - z_2)^2}. \tag{2.20}$$

This approach would treat the resource dimension $z \in \mathbb{R}_+$ exactly as the position dimensions $(x, y) \in \mathbb{R}^2$ although these have a different scale, range of values and modelling importance. Given a 2-dimension Euclidean distance between the two stations, the 3D Euclidean distance is minimum for $z_1 = z_2$, irrespective of the amount of resources available. This means, if both stations have equally scarce resources, they are exactly as close in 3D as when both have equally abundant resources. This insensitivity in absolute resource values, and the treatment of both position-related and traffic-related features in the same way, suggests that the 3D Euclidean distance is not a good candidate.

A different interesting option is to use the distance of the Poincaré half-plane model in 3-dimensions [Krioukov *et al.* 2010]. This non-Euclidean geometry is restricted in the upper half-plane for the z-dimension (resources z are positive) and it is denoted by $\mathbb{H}$. The hyperbolic distance in 3D is defined as follows

$$d_{H3}(x_1, y_1, z_1, x_2, y_2, z_3) = arcosh\left(1 + \frac{(x_1 - x_2)^2 + (y_1 - y_2)^2 + (z_1 - z_2)^2}{2z_1 z_2}\right) \tag{2.21}$$

The $arcosh(s) := \ln\left(s + \sqrt{s^2 - 1}\right)$ is simply the inverse hyperbolic cosine, which is increasing in the argument s. What is interesting about (2.21) is that the resource z is treated differently than the position (x, y). If the 2D distance is fixed, the d_{H3} is larger when the product $z_1 z_2$ is smaller, i.e. two stations both with scarce resources are far from each other. Again here, for $z_1 z_2$ product fixed, the distance is minimum when the resources are balanced, i.e. for $z_1 = z_2$.

This last property, does not allow to incorporate resource complementarity directly into the clustering mechanism. A first result showing how the extra resource dimension and the hyperbolic distance affect the clusters of static pairs and singles is published.

Publication [Álvarez-Corrales *et al.* 2017]: L. D. Álvarez-Corrales, A. Giovanidis, P. Martins and L. Decreusefond, "Wireless node cooperation with resource availability constraints", *15th Int. Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOpt)*, 2017.

2.6.1 Hyperbolic K-means

The hyperbolic distance between two stations is minimum when their resources are equal. This is not the desired design property for C-RANs. A more appropriate way to introduce the resource balance criterion in clustering, i.e. to motivate stations having more resources to collaborate with stations having less resources, is the following: Given one station $u = (x_1, y_1, z_1)$, take as candidate another station $v = (x_2, y_2, z_2)$, and consider the *centroid of the pair* having coordinates (x_c, y_c, z_c). The centroid lies in the middle of the interval connecting them $(\frac{x_1+x_2}{2}, \frac{y_1+y_2}{2})$ and has as resources the mean of the two values $\frac{z_1+z_2}{2}$.

Then, the NN of station u is the one with the minimum hyperbolic distance from the corresponding centroid, i.e.

$$w \quad = \quad \underset{v \in \Phi \setminus \{u\}}{\arg\min} \quad d_{H3}\left(x_1, y_1, z_1, \frac{x_1 + x_2}{2}, \frac{y_1 + y_2}{2}, \frac{z_1 + z_2}{2}\right), \tag{2.22}$$

and we write $u \to w$. If the u is also the NN for station w, i.e. $w \to u$, then both stations are in the same cluster, and cooperate $u \leftrightarrow w$. The advantage of this approach that uses the centroid, is that the two collaborating stations will be close to each other both in Euclidean distance and difference in resources with the centroid. This allows for cooperation between one station that has sparse and another that has abundant resources, because their centroid will have average resources, thus both will be close to this third virtual point.

The benefits of this approach are more obvious when we start considering clusters of size larger than 2 (arbitrary size). In fact, using this approach, as part of two Masters thesis, we have developed a variation of the K-means clustering algorithm, which we named "hyperbolic K-means". The algorithm works as follows.

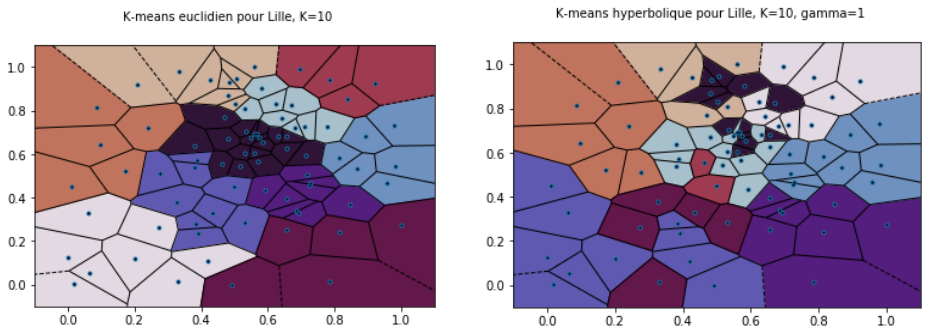

Figure 2.7: Euclidean (left) VS Hyperbolic (right) K-means for the city of Lille.

- Initialisation: Fix a number $K > 1$ of desired clusters and choose the coordinates of the K centres uniformly at random.

  1. Associate each station (x_i, y_i, z_i) with its closest centre (in d_{H3}).
  2. The centroid coordinates (cluster centres) are updated by averaging over all the stations (RRH) associated to each of them, i.e. if C_n stations are associated to some cluster at iteration n, then $\left(\frac{1}{C_n}\sum_{j=1}^{C_n} x_j, \frac{1}{C_n}\sum_{j=1}^{C_n} y_j, \frac{1}{C_n}\sum_{j=1}^{C_n} z_j\right)$ is the new centroid.

- Convergence: We iterate Steps 1. and 2. until the centres stop changing.

The approach allows resource complementarity among RRHs cooperating in a cluster. To understand this, consider the association of a single station to one of two fictional clusters. The station is slightly closer in Euclidean distance ($d_{E2-left} < d_{E2-right}$) to the left than the right cluster. But the station has resources z with value closer to the average resources of the right cluster. Then this station will be assigned to the right cluster for a large range of relative resource and distance values.

At the moment, I am in the process of applying our hyperbolic K-means static clustering algorithm to the cellular network infrastructure of cities in France, using real data for the station positions, and the traffic demand. An example of the resulting clustering for $K = 10$ clusters is given in Fig. 2.7. It shows that the Euclidean K-means produces geographically compact clusters, whereas hyperbolic K-means allows for non-neighbouring cells to be in the same cluster for resource (traffic) complementarity reasons. We are in the process of preparing with the Master students a submission presenting this new algorithm.

- **Master thesis** of Zakarya Boubazine: "Traffic-aware grouping of planar wireless nodes for C-RAN architectures", Sorbonne University, 2017.
- **Master project** of Liticia Touzari & Hanane Djeddal : "Clustering algorithms for wireless infrastructure", Sorbonne University, 2020.

2.7 Conclusions

My research on clustering and cooperative transmission of cellular stations has spanned several years and the research approach and methodology has evolved over time. I have covered both combinatorial and probabilistic approaches, whereas the last work on hyperbolic K-means departs from modelling and optimisation and introduces a new data-driven algorithm. I have worked together with 1 PhD student and several Master students for the results, which have inspired this research in several ways.

This research can continue to evolve following technological needs of the 6G, where several new IoT devices should be grouped together in an efficient way. The densification trends of future networks will accelerate, so i believe that the methods and results can evolve to adapt to new wireless network architectures and new needs. Apart from the fact that the number of nodes will become huge (so algorithms should be able to scale) another aspect that has not been sufficiently treated is clustering that supports mobility, as well as evolving cluster formations over time. A simple application is clustering of drones and other airborne access points. Such applications will be more and more important in the next years when 5G matures as the standard cellular architecture.

Chapter 3

Caching at the wireless edge

Contents

The second part of my work (2014-today) studies edge-caching architectures, where small cache memories are installed at the wireless edge. The research considers content management, user association, leasing aspects and network friendly recommendations. I supervised officially 1 PhD student, and unofficially 2 PhD students. Also, I was responsible for 1 Master thesis related to this subject.

3.1 Motivating caching at the wireless edge

In this chapter, we focus on internet traffic and content, and investigate novel ways to serve user demands, over the cellular network. It is without doubt that the prevailing traffic nowadays is HTTP and the major user requests are for multimedia, be that video or audio. Video streaming occupies the majority of internet traffic, and cellular traffic in particular, due to the proliferation of smartphone devices. This trend will not just continue in the future (horizon 2023 [Cisco 2020]) but it will become much more pronounced because of the introduction of Ultra-High-Definition (UHD), or 4K, video streaming.

Video content delivery is a major cause of congestion in the internet and a celebrated solution has been the introduction of content delivery networks (CDNs), such as the Akamai platform [Nygren *et al.* 2010], which cache popular content at

local intermediate memories ("caches") distributed over wide geographic areas. By storing popular content close to the network edge, rather than at the origin server, traffic load through the network can be reduced and bandwidth can be saved. The idea to re-direct a client to a nearby copy of locally stored content challenges todays' internet architecture. A future information centric network (ICN) envisions to equip routers with caches, and to allow content replication everywhere [Kurose 2014]. The design of an information-centric network is based upon a memory-bandwidth trade-off: how much cheaper is it to store copies of popular content close to users than to fetch them repeatedly over the Internet? [Roberts & Sbihi 2013].

Mixing these infrastructural concepts with cellular networks is also of interest, especially since most bandwidth-related benefits are harvested from caches at the network edge [Fayazbakhsh *et al.* 2013]. Starting as early as 2013, measurement studies performed on 3G traffic revealed that caching cellular content can reduce download bandwidth consumption up to 27.1% for wireless traffic [Woo *et al.* 2013]. The study in [Erman *et al.* 2011] also revealed that, caching at the cellular backhaul - ahead of the core network - can result in a hit ratio (i.e. the ratio to satisfy the demand from the cache) for the overall population of UEs of around 33%. This is a very impressive result: although cellular traffic is geographically non-homogenous, caching close to the edge offers tremendous economy in resources.

We have already seen in the previous chapter that 4G and 5G cellular networks have aimed to increase wireless throughput (area spectral efficiency) by densifying the cellular nodes, through installation of numerous additional nodes of heterogeneous size and serving capacity (micro-, nano-, pico- cells). The new architecture wants to shrink the wireless cell and bring the user as close as possible to the cellular transceiver, thus offering a strong signal over the whole wireless spectrum. As density increases however, the backhaul capacity connecting these nodes to the internet becomes the system bottleneck. So, why not extend the CDN idea to the wireless edge?

Caching on cellular equipment has been proposed in [Shanmugam *et al.* 2013]. The authors suggested that all the cellular nodes (and especially small cells) can be augmented by cache memory, to locally store multimedia, and especially video content. This way, highly predictable traffic can be handled directly from the edge nodes, without overburdening the backhaul and core network. This is not just a simple CDN variation. To design such a challenging architecture one needs to consider several aspects [Paschos *et al.* 2016]: where to install cache memories and how to dimension them, how to manage such storage space, and how to adapt many traditional wireless problems such as load balancing. Furthermore, very important issues are raised related to the ownership and management of such storage space: will it be the mobile network operator (MNO) or rather the content provider (CP) responsible? The example of CDNs shows that content providers can introduce and control their own storage devices on routers (see Netflix's Open Connect [Böttger *et al.* 2018]). Should the same solution be applied to wireless edge caching? Or is there a benefit for the MNO to build and maintain its own storage space, and lease it to various content providers? A serious obstacle to such viewpoint is end-to-

end encryption (HTTPS) required from all European operators, leaving little space for knowledge over the content of user requests to others but the CPs.

There are several benefits from caching at the wireless edge. The main one is, as mentioned above, traffic reduction on the backhaul and core network. Multiple requests for the same content can be served from the edge cache while the delivery from the source to the cache only needs to occur once. Additional benefits include reduced delay of service delivery, by placing content closer to the user, thus offering improved QoS. Also, cell located caches serve a relatively small area, thus traffic is spatially fine grained and stored content can adapt better to local demands. Finally, the network operator can better adapt modulation and coding of streamed video quality to varying wireless channel conditions [Seetharam *et al.* 2015] (as in DASH streaming protocol).

The question where to install cache memories is not a trivial one. As mentioned, CDNs cache content deeper in the network, e.g. on routers [Rosensweig *et al.* 2010], and apply hierarchical caching structures [Borst *et al.* 2010]. The advantage of deeper caching is that users associated to several BSs can be routed to the cache and make use of the cached content. This solution, however, does not solve the congestion at the wireless backhaul.

- **Cached Base Stations (CBSs)**: It is more profitable to install cache memories at the BBU of the base stations [Bastug *et al.* 2015], or the BBU pool of C-RANs. Similarly, aside macro base stations, small and nano cells can adopt the solution. The drawback here is that traffic in small cell coverage areas is difficult to predict and conventional user association does not take caching decisions into account. The approach can be extended to WIFI access points [Jaffrès-Runser & Jakllari 2018].

- **Mobile user devices (Ds)**: Content can be cached on smartphones or other portable user devices. Device-to-Device (D2D) content transmission allows for direct exchange of content between users (like wireless peer-to-peer) without use of the cell backhaul connections [Golrezaei *et al.* 2013]. For this solution, one needs to take into account that device caches are generally comparatively small, whereas D2D file transfer depends on the relatively weak signal emitted by smartphones and connections are unstable, due to mobility of both the sending and the receiving device. Note that devices can be a fleet of vehicles [Baron *et al.* 2016].

The main challenges of wireless edge caching that I have dealt with are:

1. Introduction of novel cache management policies adapted to wireless caching.

2. Evaluation of the role of mobility in D2D communications with caches.

3. Optimal solution of the joint cache placement and user association problem. To solve this, user association and load balancing become cache-aware.

4. Suggestion of an economic collaboration through leasing cache memory space from the MNO to the CP.

• My current research extends the results on caching, to deal with the problem of network friendly recommendations (e.g. by YouTube or some other application) to promote cached content and improve cache-related network benefits.

As **performance metric** I mostly use the *hit probability* (i.e. the probability that a demand finds the requested content inside the serving cache). Other metrics can include network delay, as in [Shanmugam *et al.* 2013]. Finally, I introduce and analyse a novel cache performance metric, namely the *service success probability*.

3.1.1 Cache Management and Content Replacement

A cache is useful when it stores content which is likely to be downloaded. The policy which decides what file is cached and when the cache inventory is refreshed is called *replacement policy*. There are two types of management policies that differ in the frequency of updates and traffic information availability.

- **Online Caching**: such policies update the cache inventory on each request. One item is inserted and one item is removed per content request. This way, the policies (try to) adapt to frequency changes of requested content and make the best use of bounded cache space, while being agnostic to actual request statistics. The family includes historically famous policies, such as Least-Recently-Used (LRU). The latter serves a request directly from the cache whenever it is available. If the request does not find the content cached, the item is fetched from the origin server and saved at the first position of the list. Since memory is finite the least-recently-used item in the inventory (last in the list) is removed. Other policies include [Garetto *et al.* 2016] Least-Frequently-Used removal (LFU) where items are ordered by frequency, First-In-First-Out (FIFO), or k-LRU where an item is inserted in the cache after $k > 1$ requests. An important generalisation includes policies where each item has a Time-To-Live (TTL), and is removed as soon as its counter expires.

- **Prefetching**: such policies utilise knowledge over traffic request statistics (mean volume per content, popularity) and cache the most popular content [Tatarinov *et al.* 1997]. The popularity of the content is either measured within a time-window by analysing log files of past requests, or it can be predicted for future requests (as in proactive caching suggested by [Bastug *et al.* 2014]). The update of cache inventory is not dynamic, in contrast to online policies. Rather the cache inventory remains unchanged for a determined period and the whole inventory is refreshed periodically (say per day or per week, as in Amazon Prime weekly top-of-the-chart titles), usually at off-peak hours. This approach is interesting especially for wireless edge-caching, because of the backhaul, which for small cells can be wireless. The prefetching can be done at off-peak hours, when resources are more available for system maintenance.

Prefetching policies that pre-fill a set of caches with content can be *deterministic* (i.e. specific content to each cache node) or *randomised* (i.e. the cached content in all nodes is randomly placed following some distribution).

3.1.2 D2D Communications and Device Mobility

To offload cellular traffic it has been suggested that user devices inside a cell can communicate with each other through direct opportunistic links without passing through the base station [Doppler *et al.* 2009]. So called device-to-device (D2D) communications can reduce the downlink traffic in cellular networks. This can be done in a more successful way when user devices are also equipped with cache memories to store popular multimedia files. A similar suggestion is to equip conventional vehicles with caches that transfer data around a city [Baron *et al.* 2016]. Such approaches have small cache deployment cost, whereas capacity scales-up naturally as the density of such devices increases. The main drawback is the randomness and fleeting nature of user encounter. In fact users are mobile, so as mobility increases the throughput and successful completion of transmission of entire files (or chunks) between users decreases [Alfano *et al.* 2016]. When users move around quickly, then the authors in [Golrezaei *et al.* 2013] suggests that the caching needs to be done randomly. In other words, each user terminal will cache files according to a probability density function (pdf).

3.1.3 User Association and Load Balancing

The conventional user association policy serves each cellular user from the station with the strongest signal. By considering long-term channel quality (with fast-fading channel fluctuations averaged-out) this is equivalent to associating the user device to its geographically closest access point (CLOSEST). However, due to the installation of numerous small base stations, a user device can be covered by several stations with signal sufficiently strong to enable wireless communications. So, in case of multi-coverage the device could choose some other potential station to connect to, at the cost of channel quality (hence throughput loss). This is already done today for reasons of load balancing, especially for users that are at the cell edge. It can be preferable to serve users with weaker signal in order to achieve fairness of resource consumption among stations.

Another reason to do so can be due to caching. If a user has the possibility to choose service among several access points, then, he can connect to the station that stores the requested content, rather than the station with the strongest signal. This is indeed a game-changing approach: in the CLOSEST association, the user is restricted to look into the cache of only one station and cannot profit from neighbouring nodes. In so called CACHE-AWARE association, the user can profit from neighbouring content. This can incentivise the cache manager to diversify content placement between neighbouring caches. Now, it becomes sub-optimal to cache the most popular content everywhere, because there is room for improvement of the

global cache hit probability. Additionally, by sending users to stations storing the requested content, the load is balanced naturally among nodes.

All papers that formulate and try to solve versions of the joint user association and caching problem, silently assume the non-conventional CACHE-AWARE association. We refer the reader to [Shanmugam *et al.* 2013], [Dehghan *et al.* 2017] and [Poularakis *et al.* 2014]. An extension is considered in [Tuholukova *et al.* 2017], where stations equipped with caches can cooperatively serve a user by performing JT CoMP, as in (2.3). In the latter case, the user is not associated to just a unique station but rather to a cooperating group.

3.2 Modelling ingredients

3.2.1 Traffic

When analysing performance of networks with caches, it is important to appropriately model request traffic. Models are usually based on more or less accurate prediction of content demand. The requests are dynamic over time, but the underlying time-series can be stationary or not.

- **Independent Reference model (IRM)**. A very famous traffic model in the caching literature is IRM. The model considers a fixed content library with cardinality F files. Requests arrive following an independent Poisson process of rate $\lambda > 0$ and a request is for content j with probability a_j. In fact, this probability is the *popularity* of the file. The model assumes that the probability mass function (pmf) $\{a_j\}_{j=1}^{F}$ is *static* - which models real traffic only approximately and should be restricted to finite time-windows. Popularity is Zipf distributed [Newman 2005]. Zipf distribution considers an order from the most popular file with index $j = 1$ to the least popular with index $j = F$. The probability that a request is for content j is equal to

$$a_j \quad = \quad C^{-1} \cdot j^{-\gamma} \tag{3.1}$$

  The Zipf exponent γ, is chosen often within the range $\gamma \in (0, 2]$, to limit the difference between the two most popular files, i.e. $a_1/a_2 \leq 4$, but it generally depends on measurements over popularity statistics. The normalisation factor A is found by summing the mass to unity

$$\sum_{j=1}^{F} a_j \quad = \quad 1. \tag{3.2}$$

  A value γ approaching zero results in a uniform popularity distribution, whereas larger values of γ produce distributions with increasingly lighter tails. So, the IRM has two tuning parameters, namely λ to control the rate of requests, and γ to control the skewness of the popularity. The authors in [Breslau *et al.* 1999] show that request distributions from web-proxy caches are almost Zipf distributed.

- **Time-varying models**. The IRM assumptions over static catalog and independent requests is far from reality. Real content in the web is ephemeral with rapidly changing popularity. A content appears at some point, it becomes increasingly popular and then interest fades-out gradually. The authors in [Leonardi & Torrisi 2015] have introduced a model for traffic exhibiting temporal locality, which they call the *shot noise model*. In such model the catalog evolves over time. New content arrives with rate λ_c. After the moment of appearance, say at t_c, the content is related to a volume of requests and a life-span, summarised by a set of inter-arrival times Z_c. These form a non-homogenous Poisson process, and the total traffic is the superposition of all such processes for all arriving content.

- **File-size**. A very common modelling assumption in the caching literature is that files are considered of equal size in bits. This assumption is mainly used for computational convenience, and the authors give as practical argument that long files are often broken down to equal size chunks (blocks) of the same length. This is definitely true especially for video files, which during streaming sessions are downloaded chunk-wise, to better adapt to volatile channel conditions and to save bandwidth in case a user drops viewing the content in the middle. Netflix and Youtube, for example, partition their videos into partial files of down to 2 seconds each. However, file-size is crucial for caching applications, and the statistics of multimedia length should be taken into account. Potentially the audience is interested in the entire video not only its first chunk. We find in the literature, various papers with measurements over the file-size distribution of video on the web content. Most of these find heavy-tailed distributions, and specifically: Pareto in [Crovella & Bestavros 1997], Weibull in [Abhari & Soraya 2009] and [Lee *et al.* 2013], log-normal in [Downey 2001].

Additionally to temporal locality, real traffic exhibits also spatial locality [Traverso *et al.* 2015], with different geographical regions showing interest to different content.

3.2.2 Node positions and cell coverage

When we refer to nodes we mean both user terminals and base stations. We will consider models with node positions that are either *deterministic* but arbitrary, or *randomised* in 2D, as in the C-RAN chapter. The two possible models, one with placement of stations in a grid and another with placement of stations uniformly at random in a square area, are shown in Fig. 3.1. In the figure, the density of stations per square meter is the same, and the user positions are random. Random positions are modelled using homogeneous spatial Poisson processes. The process (set) for users is Φ_u with density λ_u $[user/m^2]$ and the process for stations is Φ_b with density λ_b $[station/m^2]$. The *typical user* is placed at the centre $(0, 0)$.

Two types of areas are related with each base station,

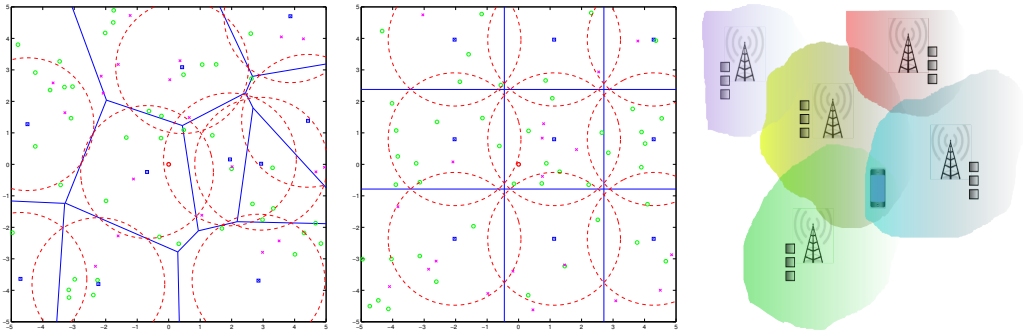

Figure 3.1: Random and grid position of base stations, with their coverage areas. Discs for SNR coverage area, asymmetric shapes for the $SINR$ coverage area.

- its *Voronoi cell*, which is the planar area closest to this station than to any other. Any two Voronoi cells are mutually disjoint.
- its *coverage cell*, which is the area covered by the station, so that any user inside can be sufficiently served by this station. Here we need to make the distinction between the SNR and the $SINR$ cell.

  1. The SNR cell is the set of points around a station, for which the received downlink SNR is larger than a QoS threshold T. We write

  $$SNR(r) = \frac{phr^{-\beta}}{w\sigma^2} > T \quad \Rightarrow \quad r < \left(\frac{ph}{Tw\sigma^2}\right)^{\frac{1}{\beta}}. \tag{3.3}$$

  The notation is the same as in (2.2) but here we consider that service from a station at r is over a frequency band w smaller than the total bandwidth W, and interference is 0. The above inequality shows that the SNR coverage cell is a disc centred at the station.

  2. The $SINR$ cell is the set of points around a station, for which the received downlink $SINR$ is larger than a QoS threshold T. We write

  $$SINR(r) \quad = \quad \frac{phr^{-\beta}}{I_{\Phi\setminus r} + W\sigma^2} > T. \tag{3.4}$$

  The difference with SNR is that now communications takes place over the entire available bandwidth W, and interference is not negligible; it stems from all stations in the network, except service node x at $d(0,x) = r$. The shape of each cell depends on the positions of neighbouring nodes and is non-symmetric.

The coverage cells of neighbouring stations may overlap. In that case we say that *multi-coverage* occurs, and a user can be served by any of its covering stations. Their count is called the **coverage number.** The multi-coverage probability in both the SNR and the $SINR$ coverage case is defined as

$$p_m \quad = \quad \mathbb{P}\left(\mathcal{N}(T) = m\right), \quad m = 0, 1, \ldots \tag{3.5}$$

This is the probability that a user is covered by exactly m stations, when the reception threshold is T. We can write the coverage number as a function of the $S(I)NR$. For the SNR cell the $\mathcal{N}(T)$ r.v. is Poisson distributed. For the $SINR$ case, the coverage number probability mass function (pmf) is derived in [Keeler *et al.* 2013].

3.3 Randomised geographic prefetching

My first work on wireless edge-caching was published at a moment when first results in the area had just started to appear. It was objectively an influential contribution to the field which has been cited more than 300 times to date. Our paper revisited the work in [Shanmugam *et al.* 2013] which introduced Femtocaching. In their paper a binary program decides which content to cache on which stations, while the users and stations form a bipartite graph. Each station was equipped with a cache memory of size K files. The original problem is NP-hard and the authors solve it approximately. In our paper instead, we use randomised instead of binary caching, and formulate a continuous concave problem that can be solved exactly to optimality, and we propose a very fast solution algorithm. To obtain such formulation we introduced the coverage number probability (3.5) for multi-coverage, based on randomised geometries. This treats all stations equally, which is a correct assumption if stations are uniformly placed in an area and have uniform traffic. We optimised over the frequency of replicas in the whole network, which are continuous, instead of a group of binary variables. For content placement, we use a random K-size vector generator, which generalises the Bernoulli variable generator. This guarantees that each station gets exactly K files, while respecting the optimal frequencies.

Publication [Blaszczyszyn & Giovanidis 2015]: B. Blaszczyszyn and A. Giovanidis, "Optimal geographic caching in cellular networks", *IEEE International Conference on Communications (ICC)*, 2015.

Specifically, we consider a prefetching cache management policy on base stations. The content catalog is fixed with F files and the popularity of content is assumed to be known and follows some distribution $\{a_j\}_{j=1}^{F}$. Zipf can be a special case. Nodes are placed randomly and we focus on a typical user asking for content following $\{a_j\}$. The user is covered by $m = 0, \ldots, M$ stations with probability distribution $\{p_m\}_{m=1}^{M}$. All stations are equipped with a cache memory of size K slots, and all files are of equal size. We assume a randomised placement policy on all cache memories, so that the frequency of replicas for the $j = 1, \ldots, F$ files is the *unknown* distribution $\{b_j\}_{j=1}^{F}$ which we are looking for.

The following program maximises hit-probability for the typical user,

$$\begin{array}{ll} \max_{\{b_j\}} & \sum_{j=1}^{F} a_j \sum_{m=1}^{M} p_m \left(1 - (1 - b_j)^m\right) \\ s.t. & \sum_j b_j \leq K \\ & 0 \leq b_j \leq 1, \, \forall j \end{array} \tag{3.6}$$

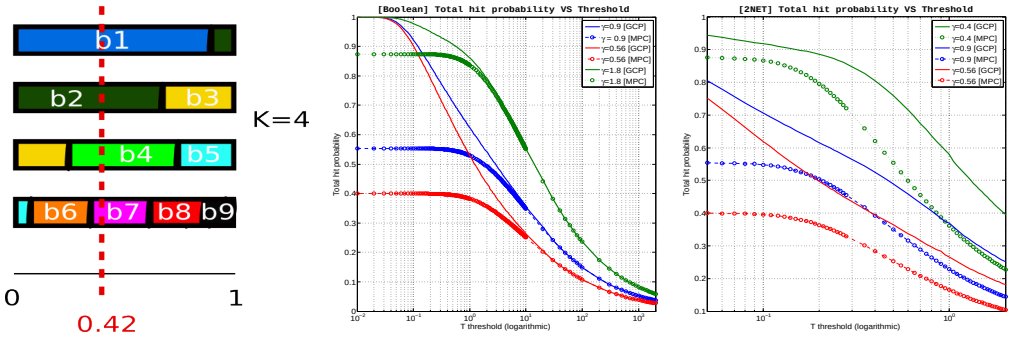

Figure 3.2: (left) Randomised vector generator, (middle) Hit-probability benefits for SNR coverage, (right) Benefits for $SINR$ coverage and two overlapping networks.

We assume that placement of content is independent per station, so that the probability that no covering station has content j is the product $(1-b_j)^m$. The problem is concave and separable in the placement frequencies b_j, so it can be solved very efficiently. The tricky part is the frequency constraint

$$\sum_j b_j \quad \leq \quad K. \tag{3.7}$$

We have shown in the paper that this constraint is necessary and sufficient so that a *placement policy* inserts content to stations with individual frequencies $\{b_j\}$, while it satisfies that at most K files are inserted per station. The placement policy to achieve this is constructed as a *random vector generator*, shown in Fig. 3.2.

The problem in (3.6) can be efficiently solved even for large values of K and M. The solution finds several objects with placement frequency less than 1 and more than 0, so we conclude that a diversified placement of content among stations – even in a randomised way – has considerable hit-probability benefits, compared to the simple strategy of just caching the K most popular files everywhere.

We have applied our solution to the case of a Zipf popularity distribution and both the SNR and $SINR$ coverage cases. The benefits of the scheme are compared to the K most popular cache placement. Benefits are pronounced for the SNR case (frequency reuse among stations). The $SINR$ case shows less benefits because, due to interference, multi-coverage is rare. The benefits are more considerable for the $SINR$ case with two overlapping networks, where the user can choose the one or the other for service (e.g. cellular stations and WIFI hot-spots), because there will always be at least 2 stations to choose from. The performance is shown in Fig. 3.2.

3.4 Spatial multi-LRU

As a next step, I investigated online cache management policies that profit from multi-coverage. This subject was very original at the time, because most online replacement policies are designed for a single cache. The standard online policy for single cache is the LRU, whose exact performance analysis is notoriously difficult

[Dan & Towsley 1990]. Using the approximation that the time spent from every item inside the cache memory is almost constant and equal for all items, Fagin [Fagin 1977] and Che [Hao Che *et al.* 2002], managed to derive a sufficiently precise way to derive the LRU performance, which was explained in [Fricker *et al.* 2012] why it works that well. The authors in [Garetto *et al.* 2016] used the approximation to study many other online replacement variations.

Our work with the Master student Apostolos Avranas was the first to propose an online replacement policy involving simultaneously several caches.

• **Master thesis** of A. Avranas: "Caching policies in wireless networks with coverage overlaps", Télécom ParisTech, 2015.

We called the strategy *multi-LRU*, as it used as basis the LRU mechanism. We studied two variations, the *multi-LRU-One* and the *multi-LRU-All*, which differ in the number of replicas inserted simultaneously in the involved caches. Furthermore, we studied the performance of the mechanism for both the IRM static traffic, as well as the traffic exhibiting temporal locality. We produced two publications,

Publication [Giovanidis & Avranas 2016]: A. Giovanidis and A. Avranas, "Spatial Multi-LRU Caching for Wireless Networks with Coverage Overlaps", *SIGMETRICS Performance Evaluation Review*, 2016.
Publication [Avranas & Giovanidis 2016]: A. Avranas and A. Giovanidis, "Performance of spatial Multi-LRU caching under traffic with temporal locality", *9th International Symposium on Turbo Codes and Iterative Information Processing (ISTC)*, 2016.

Our aim was simple: build an online replacement mechanism, completely agnostic to the traffic statistics, that can profit from *multi-coverage*. We wanted an online version of the geographic caching, but without the knowledge over the popularity statistics $\{a_j\}$. The most simple starting point was the LRU for a single cache, which works as follows:

- **[a. Update]** If the request finds the file inside the cache inventory, then the request is served from the cache and the file is moved to the first position of the inventory.
- **[b. Insertion]** If the request does not find the file inside the cache inventory, then the file is fetched from the origin data centre, and it replaces one item from the cache. The replaced item is the file at the last position, i.e. the *Least-Recently-Used* item.

The idea behind LRU is that a request for some item is also an indicator of future demand. The served item then is inserted in the cache; it moves one step towards exit every time a different object is requested, but it is moved back to first position each time a new request comes for the same item.

We extend this to the case of multi-coverage, where $m > 1$ stations cover a user. Instead of first associating the user to a unique station and then checking its cache, we assume that the user is associated to all covering stations, searches all their caches, and then is associated to the station which has the file, or to some arbitrary (e.g. closest) if none has it. In this way, the user can profit from mK cache slots, instead of just K. How are the inventories of these m memories updated? We propose two different mechanisms and find Che-like approximations:

• **multi-LRU-One**: Action is taken only in one cache out of the covering m. **[a. Update]** If the content is found in a non-empty subset of the m caches, only one cache from the subset is used for download and, for this, the content is moved to the most-recently-used position. **[b. Insertion]** If the object is not found in any cache, it is downloaded and inserted only in one cache, and its Least-Recently-Used object is evicted. This one cache can be conventionally chosen as the closest to the user.

The total hit probability is approximated as

$$P_{hit-ONE} \overset{IRM}{=} \sum_{j=1}^{F} a_j \sum_{m=0}^{M} p_m \left(1 - e^{-a_j \lambda_u m |\mathcal{V}| T_C}\right). \tag{3.8}$$

where $|\mathcal{V}|$ is the (mean) Voronoi surface and T_C the characteristic time for all objects, found by solving $\sum_{j=1}^{F}(1 - \exp(-a_j \lambda_u |\mathcal{V}| T_C)) = K$.

• **multi-LRU-All**: Insertion action is taken in all m caches. **[a. Update]** If the content is found in a non-empty subset of the m caches, all caches from this subset are updated. **[b. Insertion]** If the object is not found in any cache, then it is inserted in all m. A variation based on q-LRU can be proposed, where the object is inserted in each cache with probability $q > 0$.

The total hit probability is approximated as

$$P_{hit-ALL} \overset{IRM}{=} \sum_{j=1}^{F} a_j \sum_{m=0}^{M} p_m \left(1 - e^{-a_j \lambda_u |\mathcal{A}_m| T_C}\right), \tag{3.9}$$

where $|\mathcal{A}_m|$ is the (mean) $S(I)NR$ joint coverage surface from all the m stations. T_C is again the characteristic time for all objects, found by solving $\sum_{j=1}^{F}(1 - \exp(-a_j \lambda_u |\mathcal{C}| T_C)) = K$, where $|\mathcal{C}|$ is the surface of one coverage cell.

The motivation behind the two versions is the following. If multi-LRU-One is applied, a single replica of the missed content is left down in one of the $m > 1$ caches, thus favouring diversity among neighbouring caches. If multi-LRU-All is used, m replicas are left down, one in each cache, thus spreading the new content over a larger geographic area (the union of m covering cells), at the cost of diversity. A q-multi-LRU-All version is in-between the two, leaving down a smaller than m number of replicas.

The performance of the two versions (-One and -All) largely depends on the type of incoming traffic. For fixed object catalogue and stationary traffic (IRM), diversity in the cache inventories can be beneficial so multi-LRU-One is better (see

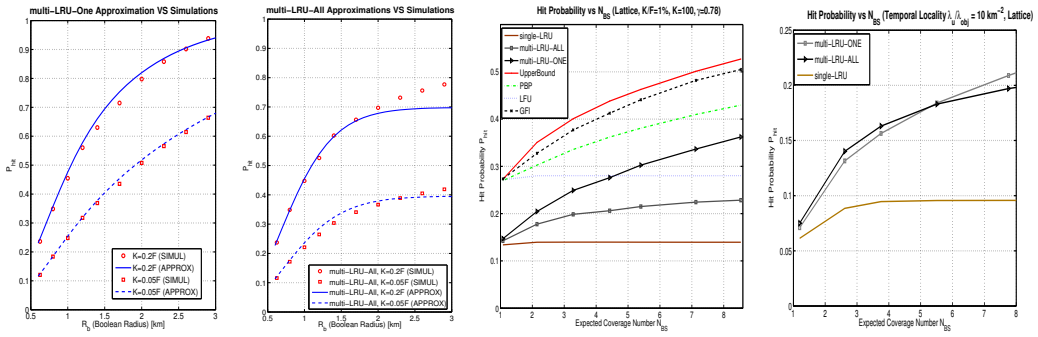

Figure 3.3: (left x2) Evaluation of the Che-approximations for -One and -All, (centre) Hit-probability comparison under IRM traffic between: multi-LRU, single-LRU, LFU, randomised geographic (PBP) and femto-caching (GFI), (right) multi-LRU-One and -All for traffic with temporal locality.

Fig.3.3,centre), whereas for time-dependent traffic with varying catalogue, performance can be improved when many replicas of the same object are available, before its popularity perishes (see Fig.3.3,right).

In [Leonardi & Neglia 2018] the authors extend the multi-LRU strategies, by proposing a new one called *lazy-LRU* which updates a cache's inventory only when it is the only one that holds the content (and serves the user).

3.5 D2D cache-hit under mobility

In this part of the work I will discuss cache-hit in D2D communications, when user terminals are equipped with cache memories to store content; direct communications between users, without passing from the base station, is enabled. In such type of communications the factor that makes performance analysis challenging is user mobility. User nodes change position continuously, thus links can be created or dropped depending on the users relative position. The work was done in collaboration with a PhD student candidate from Tunisia, who had visited Télécom ParisTech during her thesis.

- **PhD Thesis** Chedia Jarray (visitor from Tunisia, Gabes in 2015).

We started the project in 2015 and our collaboration produced two publications on this subject, as part of her PhD work.

Publication [Jarray & Giovanidis 2016]: C. Jarray and A. Giovanidis, "The effects of mobility on the hit performance of cached D2D networks", *14th International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOpt)*, 2016.
Publication [Jarray & Giovanidis 2018]: C. Jarray and A. Giovanidis, "Successful file transmission in mobile D2D networks with caches", *Elsevier Computer Networks*, 2018.

Consider a number of user-devices randomly placed on the 2D plane. We again use the spatial Poisson process distribution Φ with device density $\lambda > 0$ per m^2. At the cartesian origin, we assume a *typical device* asking for a content, with probability (popularity) that follows $\{a_j\}_{j=1}^F$. Each device has a cache memory, having files stored, and we assume a randomised content placement policy (or some online policy) which stores contents with probability $\{b_j\}_{j=1}^F$ (e.g. as in Ch.3.3). The origin device can create a link with any other device, as long as it lies inside the SNR or $SINR$ coverage cell of the other. For successful file transmission, two conditions need to hold simultaneously:

1. The associated transmitter device needs to store the requested file in its cache.
2. The connection needs to stay alive, till file transmission is completed.

In our problem we consider that every file has a given file-size $\{z_j\}_{j=1}^F$. Both conditions are fulfilled simultaneously by the associated transmitter $x \in \Phi$ if

$$\Psi_{j,x} = \mathbf{1}[j \ in \ cache \ x] \cdot \mathbf{1}\left[\tau_x \cdot w \log_2\left(1 + SNR(r_x)\right) \geq z_j\right] \quad = \quad 1, \qquad (3.10)$$

where $\mathbf{1}[\mathcal{E}]$ is the indicator function, equal to 1 when the event $\mathcal{E}$ is true. The first indicator answers whether the requested file j is in the cache of the associated transmitter x. The second indicator answers whether the total information *bits* transmitted during the time τ_x that the connection is alive are enough to complete the wireless transfer of file j with size z_j. The SNR takes the expression in (3.3), but we can also use the expression for $SINR$ in (3.4), with some frequency reuse factor. We assume that the link quality is related to throughput by the Shannon capacity formula.

Mobility: The mobility in this work is modelled in a simplified way. The link between the origin device which requests and the association device at x which transmits, stays alive during a random period of duration τ_x. When the timer expires, the two devices are considered to have moved far away from each other, so the link gets inactive. The transmitter has τ_x time to send the entire file of size z_j.

Association: Since we consider a coverage model, the association rule is important, because it determines the distance r_x between the origin and the transmitter. We investigate three cases: (a) associate to the best transmitter that can fulfil (3.10), as if an oracle knows and picks the best choice, (b) associate to the CLOSEST transmitter, irrespective of the cache inventory, (c) associate to the closest transmitter who holds a replica of the request j; this is now a CACHE-AWARE rule.

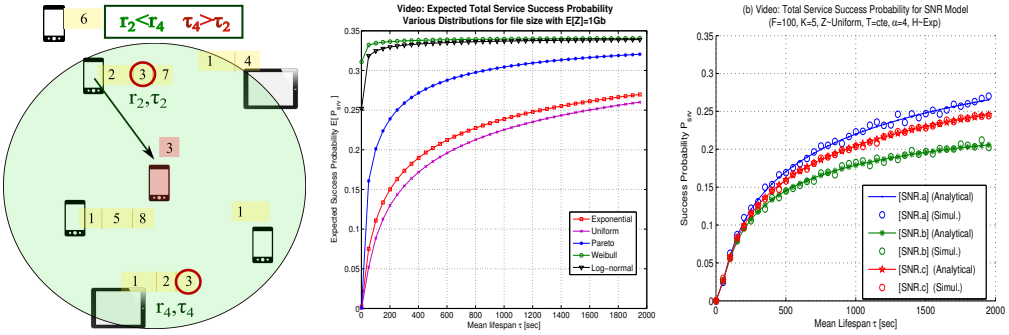

Figure 3.4: (left) D2D scenario: the closest transmitter does not have the requested file "3", the two transmitters who cache it and cover the origin differ in distance and link-lifespan (middle) Expected SSP metric for different distributions of video size, (right) Expected SSP for different association rules; link-lifespan is constant.

To evaluate performance, we introduce the so called *Service Success Probability (SSP)*, which generalises hit-probability to include the guarantee that the transmission is completed. We define,

$$P_{ssp}(\{b_j\};\{a_j\},\{z_j\}) \quad = \quad \sum_{j=1}^{F} a_j \mathbb{P}\left[\Psi_{j,x}=1\right]. \tag{3.11}$$

If we do not know the exact file-size, but rather the statistics (see Ch.3.2.1) then we can calculate the *Expected SSP* instead $\mathbb{E}[P_{ssp}(\{b_j\};\{a_j\}]$, where z follows some heavy-tailed distribution. We have derived closed-form expressions of the new metric for all association policies, and various video-size distributions, lifespan distributions, as well as fading distributions. For example, for the special case of best association (a), constant lifespan τ for all devices, Rayleigh fading and path-loss exponent $\beta=4$, we get with $g(z_j,\tau) := \sigma^{-1}\Gamma(3/2)\sqrt{p/w}\sqrt{2^{z_j/(\tau w)}-1}$

$$P_{ssp}(\{b_j\};\{a_j\},\{z_j\}) \quad = \quad \sum_{j=1}^{F} a_j \left(1-\exp\left(-b_j\lambda g(z_j,\tau)\right)\right). \tag{3.12}$$

The Expected SSP for different video-size distributions with mean size $1Gb$ is illustrated in Fig.3.4 (middle). The same metric for the three association policies (a), (b), (c) is shown in Fig.3.4 (right). CLOSEST is better than CACHE-AWARE.

Our published work has also extended the results for the case of consecutive associations, till service completion. When the service is not completed by the first transmitter, the device at the origin tries to recover the rest by some other. We have studied the performance in two variations: when only device departures occur, as well as when both departures and new device arrivals occur.

Our work is unique, because it combines file-size with service interruption due to mobility, while taking the connection strength and throughput explicitly into account.

3.6 Joint leasing, caching and user association

The original femto-caching paper [Shanmugam *et al.* 2013], considers a given set of stations and a set of users placed over some area. It allows each user to be served by an a-priori determined set of covering stations and searches for the optimal binary content placement strategy, to minimise service delay. The authors show that this problem is NP-hard and solve it with a polynomial-time approximation algorithm that gives a solution with a guaranteed distance from optimality. The authors in both [Dehghan *et al.* 2017] and [Poularakis *et al.* 2014] formulate the joint content placement and user association problem and they both show that the joint problem is NP-hard and use approximation algorithms to get a sub-optimal solution.

During the thesis of Jonatan Krolikowski, which i have supervised, we formulate the joint memory leasing, content placement and user association problem in a similar wireless setting as the above papers, with deterministic node positions and deterministic cache decisions. We include the two types of possible user association policies (CLOSEST and CACHE-AWARE).

• **Ph.D. thesis supervision** of Jonatan Krolikowski [Krolikowski 2018]: "Optimal Content Management and Dimensioning in Wireless Networks", Co-supervised by Marco Di Renzo. Université Paris-Saclay, STIC , 2018. Funded by the LABEX-Digicosme project "CONTAIN".

We present an algorithmic solution that solves the joint NP-hard problem exactly. The algorithm is guaranteed to find the global optimum, without approximation. To reassure the reader that what we claim is theoretically sound, we note here that our algorithm need not run in polynomial time; but in all tested cases the convergence was achieved after a very small number of iterations. It is the only algorithm in the literature that optimally solves the joint user association and content placement problem. As we realised later, we were not the only ones to use this method. The authors in [Bektas *et al.* 2008] have solved a much simpler caching and request routing problem using a similar approach, by restricting the formulation to linearised integer problems. Here, we formulate and solve a mixed integer non-linear program (MINLP) exactly, for the wireless setting.

Another benefit of our problem formulation is that it naturally results in a business interaction between a Mobile Network Operator (MNO) who builds cache memories on its wireless infrastructure, and associates users to stations, and a Content Provider (CP) who rents the memories and places content in them.

Specifically, we consider a planar area, with a number M of base stations at given positions. Each station is equipped with a memory of size K [*bytes*]. We investigate deterministic prefetching. There is a fixed content catalog of size F and each file has size z_f. The decision to place file f in the cache of station m is the binary variable $x_{m,f} \in \{0, 1\}$. The CP will lease part of the memory w_m and decide over the content placement. Hence the CP controls the variables $\{x_{m,f}\}$ and w_m,

where

$$\sum_f z_f x_{m,f} \leq w_m \qquad \& \qquad w_m \leq K. \tag{3.13}$$

Each station has its coverage $S(I)NR$ cell. The user traffic in the cell can be served by the covering station, but some areas are covered by more than one cell. We split the area into regions s, and each region has its own set of coverage stations $\mathcal{M}(s)$. It also has its own traffic N_s, which is the sum of requests for all files $N_{s,f}$. The traffic and popularity are known for each region. The MNO associates a portion of traffic $y_{m,s,f}$ from region s to station m about request f, and it must hold

$$\sum_{m \in \mathcal{M}(s)} y_{m,s,f} \quad \leq \quad N_{s,f}. \tag{3.14}$$

The MNO can follow the standard CLOSEST assignment or choose CACHE-AWARE. In the later case the MNO helps the CP to obtain a higher hit-rate.

The CP wants to get the maximum hit-rate, while keeping investment costs low,

$$\max_{x,y,w} \quad \sum_{m=1}^{M} U_m \left(\sum_s \sum_f y_{m,s,f} \right) - q \sum_{m=1}^{M} w_m. \tag{3.15}$$

Lets call $h(y) := \sum_{m=1}^{M} U_m \left(\sum_s \sum_f y_{m,s,f} \right)$. In the above $\sum_s \sum_f y_{m,s,f}$ is the total traffic routed to station m and $U_m(.)$ is the utility function for the traffic. The U_m allows for great flexibility in problem formulation. It can be linear when the CP is only interested in hit-rate. If the utility function is concave with diminishing returns, it can allow for load balancing among stations. The q is the price per leased memory unit. The optimisation is done over all feasible leasing, association and allocation decisions described by the constraints (3.13) and (3.14). We need to include the following coupling constraint

$$y_{m,s,f} \quad \leq \quad N_{s,f} x_{m,f}, \tag{3.16}$$

in order to count as hits, only the traffic that finds its request cached at the station.

We have produced three publications on this subject during the PhD thesis.

Publication [Krolikowski *et al.* 2017]: J. Krolikowski, A. Giovanidis and M. Di Renzo, "Fair distributed user-traffic association in cache equipped cellular networks", *15th International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOpt)*, 2017.

Publication [Krolikowski *et al.* 2018b]: J. Krolikowski, A. Giovanidis and M. Di Renzo, "Optimal Cache Leasing from a Mobile Network Operator to a Content Provider", *IEEE INFOCOM 2018 - IEEE Conference on Computer Communications*, 2018.

Publication [Krolikowski *et al.* 2018a]: J. Krolikowski, A. Giovanidis and M. Di Renzo, "A Decomposition Framework for Optimal Edge-Cache Leasing", *IEEE Journal on Selected Areas in Communications*, 2018.

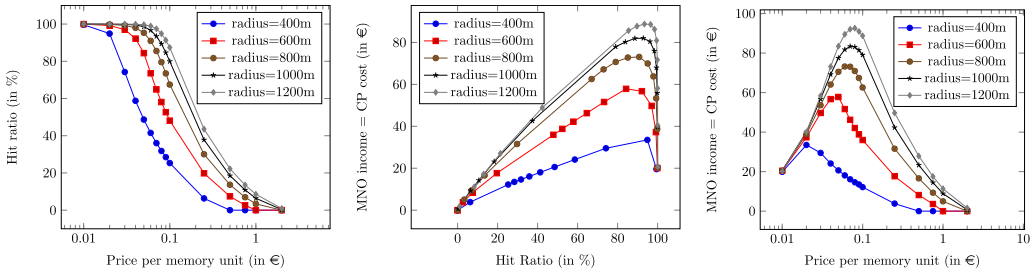

Figure 3.5: Linear utility function (hit rate), various coverage radii: (left) Hit rate achieved for price q per memory unit, (middle) total CP investment for a target hit rate, (right) optimal investment budget for price q per memory unit.

The exact optimal solution is found by applying Benders decomposition to the above NLMIP. The procedure is a Master-Slave decomposition with row generation. The main problem in (3.15) is split in two sub-problems

$$\begin{aligned} (Master) \qquad & \max_{x,w} \; h(y(x)) - q \sum_{m=1}^{M} w_m \\ (Slave) \qquad & y(x) = \arg\max_{y} \; h(y) \end{aligned}$$

The Master is solved by the CP, using the user association solution y as known. The Slave is solved by the MNO using the leasing and content placement decisions x, w as known. Iteratively, first Slave is solved and produces a new linear feasibility inequality at iteration $t = 1, 2, \ldots$

$$h(y(x)) \leq h(y^{(t)}) + \sum_{m,s,f} \lambda^{(t)}_{m,s,f} \left(N_{s,f} x_{m,f} - y^{(t)}_{m,s,f} \right), \tag{3.17}$$

where $\lambda^{(t)}_{m,s,f}$ are the Lagrange duals. Then this cut is included in the Master, which is solved with a finer constraint set. As t increases the iteration is guaranteed to converge to the optimal solution. Note that the Master is linear integer and can be solved by Branch-and-Bound with a custom solver like CPLEX. The Slave is continuous linear or concave and can be solved by Linear or Convex Programming. We have developed a distributed solution with minimum message passing among stations that solves efficiently the general assignment Slave problem.

The optimal solution for linear utilities U_m is shown in Fig. 3.5. Such utility leads to hit-rate maximisation minus investment cost. In Fig. 3.5 (left) we see that the higher the price q set by the MNO, the less the CP will invest. There is an optimal price for the MNO that can lead the CP to maximise its total investment, see Fig. 3.5 (right). The price and benefits depend on the extend of multi-coverage! This research has shown that the MNO and the CP can collaborate in a mutually profiting way, to increase hit rate at the edge. There is a profit margin for the MNO as long as it sets an appropriate price and does CACHE-AWARE association.

3.7 Network Friendly Recommendations

In early 2018, I had discussions with T. Spyropoulos and his PhD student T. Giannakas from EURECOM, France, on a very interesting extension of caching. For most caching systems the cache-hit depends on the exogenous user-demand for content, which is not easily predictable and cannot be controlled. On the other hand, most multimedia applications like YouTube, Netflix or Spotify include a recommendation engine to propose content to users, thus shaping user demand. Their idea was to develop sequential recommendation policies over long user viewing sessions, which should suggest relevant content to the users while at the same time favour cached content from the network. This way hit rate could be increased by shaping user demand, keeping some quality guarantees. This research is part of the thesis,

- **Ph.D. thesis** of Theodoros Giannakas (2018-2020, EURECOM, France), involvement and joint publication with his supervisor T. Spyropoulos

where I contribute in one journal and one conference submission:

Submission [Giannakas *et al.* 2020]: T. Giannakas, A. Giovanidis and T. Spyropoulos, "MDP-based Network Friendly Recommendations", 2020, with e-Link: [under review]. Also, conference version submitted.

My involvement consists in a Markov Decision Process (MDP) formulation of the problem, which is an area where I have concrete experience from my own PhD thesis. Indeed, the Network Friendly Recommendations (NFR) over long-sessions can be naturally cast in the framework of MDPs (see [Choi *et al.* 2019] for MDPs related to video delivery).

Suppose the length of viewing session is a random variable L geometrically distributed with mean $\overline{L} = (1-\lambda)^{-1}$, and $\lambda > 0$. The mean length can vary from 1 to ∞. In each of the sequential L steps, the user views a multimedia content and receives a batch of $N \geq 1$ other files, as recommendation from the system. The user may either choose among the N recommended files for the next viewed content, or reject all these options and search on his own some other file from the total content library of size $K > N$. If the user searches on his own, we assume the user preference K-vector is known, denoted by p_0. After the session length finishes the user exits.

The standard baseline policy is the one that suggests to the user the top-N quality items. These are the files which have most relevance to the currently viewed item $S_t = i$; we assume here that the relevance relations of items is summarised in a $K \times K$ matrix U, which is known, where $u_{ij} \geq 0$ is the weight of relation between the two files. In this work, we assume a subset M of the K catalog files as *cached*. Given this, we aim to deviate from the $top-N$ policy, and find a new optimal recommendation policy R^* that maximises *cache-hit-rate* over the whole session,

under some quality guarantees $\mathcal{R}$,

$$\max_{R\in\mathcal{R}} \quad \mathbb{E}\left[\frac{1}{L}\sum_{t=1}^{L} c(S_t)\right]. \tag{3.18}$$

In the above, S_t is the state of the system at some time $t = 1, \ldots, L$. The state is the currently viewed file, chosen from the content library of size K. The gain $c(S_t)$ is related to whether the specific viewed file is cached or not, so

$$c(S_t) \quad = \quad \begin{cases} 1, & \text{if } S_t \text{ is cached,} \\ 0, & \text{if } S_t \text{ is not cached} \end{cases}. \tag{3.19}$$

The random variables are the session length L and the $\{S_t\}_{t=1}^{L}$ vector of visited states by the user during the session, due to random user behaviour.

The difficulty often encountered when solving MDP problems is the so-called curse of dimensionality. In our case this curse appears if one uses as action set the ensemble of possible N-sized recommended batches per viewed content. As an example, for a library of size $K = 1000$ files and a batch of size $N = 3$ recommendations, there are $\frac{K!}{N!(K-N)!} \approx 166$ *Million* possible batches to chose from. Such problems cannot be solved for large libraries. To overcome this issue, we need to avoid the search over the optimal N-sized batch. Instead, we introduce $K \times K$ unknowns, which represent the *per-item recommendation frequencies*. For each viewed content $i = 1, \ldots, K$, we introduce a K-sized vector $\mathbf{r}_i$ of recommendation frequencies, whose j-th element represents the frequency that file j is included in the recommendation batch, when item i is viewed. The optimal policy $R^* = [\mathbf{r}_1^*, \ldots, \mathbf{r}_K^*]$ is the $K \times K$ matrix of such frequencies. When the user views content i, and the optimal frequency vector $\mathbf{r}_i^*$ is known, the optimal N-sized batch can be sampled (drawn) in a way similar to the random vector generator in Fig. 3.2 (left). Hence, R^* can sufficiently determine the optimal batch.

The transition probability that the user will view next file j, given that the user currently views file i is defined as

$$P_{ij} \quad = \quad \alpha_{ij} r_{ij} + \left(1 - \sum_{k=1}^{K} \alpha_{ik} r_{ik}\right) p_0(j). \tag{3.20}$$

To understand this, suppose the recommendation policy is deterministic and a single batch w is always suggested when file i is viewed. Then, $r_{ij} = 1$ for the N recommended files and $r_{ik} = 0$ for the rest. From the above $P_{ij} = \alpha_{ij}\mathbf{1}_{j\in w} + \left(1 - \sum_{k\in w} a_{ik}\right) p_0(j)$. In simple words, the user prefers recommended object j with probability α_{ij}. With probability $\left(1 - \sum_{k\in w} \alpha_{ik}\right)$ the user rejects all recommendations in the batch and chooses file j from the search bar with probability $p_0(j)$ based on personal preference. The recommendation frequency vector $\mathbf{r}_i$ needs to satisfy a set of constraints $\mathcal{R}_i$, such that the recommended batch cannot offer quality worse than some percentage of the baseline quality from the top-N related items. We find the optimal solution using Bellman equations for the discounted MDP with discount factor λ. The solution is derived iteratively by Policy

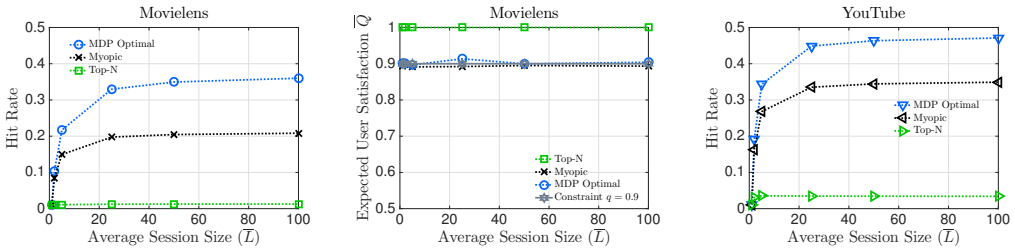

Figure 3.6: Comparison of performance for recommendation policies over two datasets. *MovieLens:* (left) Hit-Rate, (centre) Quality. *YouTube:* (right) Quality.

Iteration. Each Bellman equation can be solved by maximising over the K-vector $\mathbf{r}_i$ of unknown frequencies, which form a small LP. Thus, the inner loop has low complexity and can be parallelised.

In Fig. 3.6 we show the hit-rate and quality performance of our optimal MDP-policy, compared to the myopic-optimal and the $top-N$ baseline. The myopic-optimal finds the batch which is optimal just for the next step, without considering the length L of the user session. The datasets about file relations U are taken from a MovieLens and a YouTube multimedia database. Regarding MovieLens, we observe from Fig. 3.6 (centre) that the $top-N$ has maximum recommendation quality, but from Fig. 3.6 (left) it has lowest cache-hit-rate. The MDP has maximum hit-rate compared to both the myopic and the $top-N$ (left), while keeping the quality above the defined threshold (see constraint $q = 0.9$ (centre)). The MDP policy benefits from the offered flexibility in quality constraints to recommend cached content to the user. The benefits of the MDP policy are considerable compared to the myopic for both datasets (see Fig. 3.6 (left) and (right)), as the session-length increases.

The MDP optimal policy with item-frequencies is solved in reasonable run-time (about half an hour for $K = 10000$ and $N = 3$ in a small 8Gbyte RAM laptop).

3.8 Conclusions

In this chapter, I have presented several original results related to cache management policies in wireless cellular networks. One common aspect binding most of this work is the focus on edge-caching. With edge-caching the wireless edge inherits advanced functionality. It is not restricted anymore to signal reception and scheduling, but now can store and forward multimedia content. A fresh development of edge functionality in the new era of communication networks is the *intelligent edge*. With the advancement in artificial intelligent and the proliferation of Internet-of-Things devices in the network, the wireless edge will be required to serve not only as a memory space, but also as a local computing space, where computational burden for simpler tasks can be offloaded. The main lines of such visionary network operation have been drawn by the authors in [Park *et al.* 2019], who explain how the novel computational challenges of the machine learning (ML) era will affect the network. Specifically, two types of novelties are expected to be integrated: ML

for communications, where tools from ML can be used to adapt the system more appropriately to ever-changing channel conditions; Communication for ML, where the network needs to locally collect data and perform machine learning tasks that should cover global needs. This fragmented data collection and ML operation gives rise to a new challenge: *federated learning for wireless*. This is one major area of research that I am planning to explore next.

What is more, the network friendly recommender in Ch. 3.7 can better adapt caching and forwarding decisions to human preferences. Our solution was an MDP-based policy, which assumes a specific (but rather general) human behaviour to be known and fixed. This is not true in the real world, so a natural extension of our work is the adaptation of our solution with tools from Reinforcement Learning, so that the recommendations can adapt to various types of actual time-evolving human requests.

Chapter 4

Post diffusion in social platforms

Contents

The third part of my work (2018-today) develops dynamic models for online social platforms. These models use as input the social graph structure as well as posting and reposting activity, and give as output the influence of each user over any other inside the social network. The results are applied on real data traces to rank users based on their influence. The research was done in collaboration with 1 Master student. Currently 2 new PhD students are recruited to follow-up.

4.1 Social networks and existing models

The worldwide expansion of sufficient internet connectivity and the proliferation of internet users in the early 2000's prepared the ground for social networks to gradually appear and reign. These virtual networks are actually web platforms, which allow users to connect with friends, family members or even create new relations with people that share similar habits. Within a decade's time, online social platforms (OSPs) evolved into entities that play a major role in the way individuals communicate, share news and get informed.

Although OSPs differ from one another, most of them share a common structure, which allows users to post messages on their Wall and read posts of others on a separate Newsfeed. Most OSPs also permit re-posting from Newsfeed to Wall, in order to facilitate information diffusion. With each re-post (or "share", or "re-tweet") the information becomes visible to a new audience, which may choose to adopt it or not, thus spreading further the post or halting its diffusion. In this way, posts originally generated by some user circulate inside the social network.

Understanding how posts are diffused inside an OSP is of great importance. On the one hand, there can be specific individuals who would like to take advantage of the rumour spreading process inside a platform and promote malicious or fake content. On the other hand, companies would like to promote their products to as many users as possible for their advertising needs [Kempe *et al.* 2003]. Since such platforms can be the ground for both advertising and political opinion formation,

the way these operate should not be biased towards or against certain opinions; being web-based, they are vulnerable to software attacks which generate artificial users (bots, sibyls) with the aim to spread misinformation.

The topic of social platforms has generated a lot of research, focusing on three directions: (a) understanding opinion formation, (b) understanding post diffusion and "virality", and (c) ranking online social users based on their influence. The main research with analytic flair, has focused on three well-known models from the 70's to describe how a single opinion is spread gradually through the network of users, namely the Voter model [Holley & Liggett 1975], the DeGroot model [DeGroot 1974] and the SI(R) model [Newman 2010, Ch.17]. We should also add here the PageRank algorithm [Brin & Page 1998] that ranks web-pages based on their frequency of visit as measure of importance. Each one of these models gives its own explanation about post diffusion and can rank social users; but these answers are not satisfactory, because they do not take into account the specific mechanisms of an online social platform. Such models being *graph-based*, they neglect the dynamic user activity of posting and re-posting. Neither do they consider the role of the platform itself, which affects the process through the implemented Newsfeed algorithm. The latter determines which content to present and in what order, while leaving other content out of the feed, as less relevant.

4.2 A new dynamic model for OSPs

It has already been verified through data analysis, that high influence in social networks depends not only on the user's position on the graph (e.g. the number of followers) [Cha *et al.* 2010]. The user also needs to post sufficiently often, whereas his followers (and the followers of his followers) need to frequently share his posts. Based on these observations, I decided to develop a new dynamic model that borrows ideas from telecom networks, to describe the spread of influence inside a social platform. I have collaborated with two colleagues from Sorbonne University - LIP6 to work on this topic, specifically Bruno Baynat and Clémence Magnien. I first obtained local funding from the university to supervise a Master thesis in 2018 and 2019 on this subject.

- **Master thesis** of Antoine Vendeville: "How can the frequency of Newsfeed posts improve influence in a social network", Sorbonne-LIP6, 2018-2019.

The new model is inspired from internet delay analysis. Specifically, we follow the steps of Kleinrock, who used Markov models and queuing to evaluate the performance of packet-switched networks and their protocols. Instead of packets we have here posts and instead of queues we have lists; but the analysis is not straightforward... We consider a graph with N user-nodes, where the directed edges indicate a user-follower relation. A post can pass from a user to his followers. Each node is equipped with two lists: one list of size K for the user Wall, and one list of size M for the user Newsfeed. These lists describe the inventory of most recent posts that the

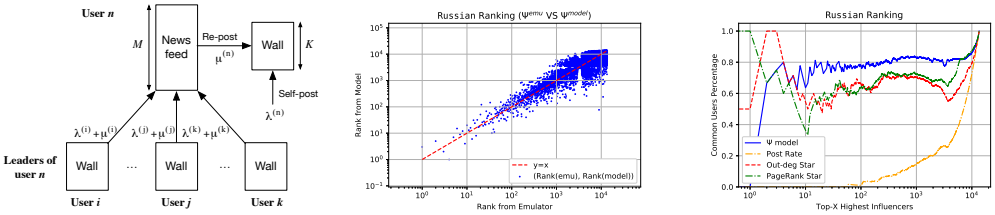

Figure 4.1: (left) The OSP markovian model: illustration of node n with Wall and Newsfeed. (centre) Comparison of user ranking between our model and the empirical influence from the Twitter trace; users are ordered from No.1 to $14K$ in decreasing order of influence. (right) Explanation of empirical ranking through our new model, compared to ranking from PageRank, posting rate, and number of followers.

user is viewing at some time instant. The lists change dynamically when new posts arrive. Each user i is related with two activity processes: one process for posting his own posts with rate $\lambda^{(i)}$ [posts per unit of time], and one process for sharing posts from his Newsfeed to his Wall with rate $\mu^{(i)}$ [posts per unit of time]. These processes are considered Poisson for the analysis, but alternative distributions can also be included. A post has origin i if it is produced and injected in the platform by user i. Fig. 4.1 (left) shows one node of this network (specifically node n).

Hence, we view the social platform as a dynamic network of (Newsfeed and Wall) lists. This is similar to the queuing network of Jackson. But this is where the similarities end. In the social platform, each time a user posts his own post or shares one post from the Newsfeed, this post enters his own Wall by replacing an older entry; at the same time this post also appears in the Newsfeeds of all followers of this user, and removes one entry. As a result, one post is cloned several times, being reproduced in all the followers' Newsfeeds. There are two policies that play a major role in post diffusion: (i) The *post eviction policy*, which is platform-dependent. Let us chose to evict at RANDOM one post from the corresponding list (Newsfeed or Wall), but FIFO or other policies can apply as well. (ii) The *post selection policy*, which is user-dependent. Based on this a user decides what post to share from the Newsfeed on his Wall. Let us chose to select also at RANDOM, but other policies can be included, e.g. always choose the first post in the list order.

We focus on some user-origin of posts i and we are looking for the unknown influence variables $p_i^{(n)}$ and $q_i^{(n)}$, for $n = 1, \ldots, N$. The $q_i^{(n)}$ is the expected percentage of posts of origin i found on the Wall of user n. They obviously satisfy for each Wall n: $\sum_{i=1}^{N} q_i^{(n)} = 1$. We can also interpret $q_i^{(n)}$ as the probability that, when picking at random a post from Wall n, this post is of origin i. Similarly, $p_i^{(n)}$ is the expected percentage of posts of origin i found on the Newsfeed of user n. These performance quantities will be the output of the analysis.

We have managed to find the exact solution to this problem. For the posts of origin i there is a set of *balance equations* on the Newsfeeds and Walls of all users in the graph. For any user $n \neq i$ we have the following equations, one for his Newsfeed

and one for his Wall.

The *Newsfeed balance equation* for the case $i \neq n$, is

$$\left[\sum_{k\in\mathcal{L}^{(n)}}\left(\lambda^{(k)}+\mu^{(k)}\right)\right]p_i^{(n)} \;=\; \lambda^{(i)}\mathbf{1}_{\{i\in\mathcal{L}^{(n)}\}}+\sum_{k\in\mathcal{L}^{(n)}}\mu^{(k)}p_i^{(k)}, \tag{4.1}$$

where $\mathcal{L}^{(n)}$ is the set of leaders of user n. The *Wall balance equation* is

$$\left(\lambda^{(n)}+\mu^{(n)}\right)q_i^{(n)} \;=\; \mu^{(n)}p_i^{(n)}. \tag{4.2}$$

Similarly, we get a pair of equations for the user $i = n$, with small differences. This gives a set of $2N$ equalities related to i, that is N for the Newsfeeds and N for the Walls. Furthermore, we consider all origins $i = 1, \ldots, N$ thus obtaining the influence of any user on any other. Altogether, this gives $2N^2$ equalities.

These equations have been derived using two different methods. One version is through a *mean field approximation*, where we focus on one Newsfeed and consider all the others to operate in steady-state. The second method is exact (no approximation involved), and applies the *conservation law of posts of origin i* in the Newsfeed (and Wall) of user n.

Interestingly, equations (4.1)-(4.2) and their counterparts for $n = i$ (not shown) allow for a simple intuitive interpretation: they balance the incoming and outgoing flow of posts of origin i on each Newsfeed and Wall list. More precisely, equation (4.1) equalizes the incoming rate and the outgoing rate of posts of origin i in the Newsfeed of user n for $n \neq i$. Here, $\sum_{k\in\mathcal{L}^{(n)}}(\lambda^{(k)}+\mu^{(k)})$ is the average number of posts per unit of time that enter the Newsfeed of user n. From the RANDOM eviction policy, each of these arriving posts replaces a post of origin i with probability $p_i^{(n)}$. Indeed, by assuming that post and re-post processes are Poisson, the PASTA property holds which tells us that arriving posts see the Newsfeed in steady-state. As a result, the left-hand side of equation (4.1) is just the outgoing rate of posts of origin i in the Newsfeed of user n. Now looking at the right-hand side of this equation, $\mu^{(k)}$ is the average number of posts per unit of time that arrive on the Newsfeed of user n because a leader k of n reposts something on his Wall. Each of these posts is of origin i with probability $p_i^{(k)}$, due to the RANDOM selection policy in Newsfeeds. In addition, if i is a leader of n, the $\lambda^{(i)}$ self-posts of i per unit of time also appear on the Newsfeed of n. As a result, the right-hand side of equation (4.1) is the incoming rate of posts of origin i in the Newsfeed of user n.

Equation (4.2) equalizes the incoming rate and the outgoing rate of posts of origin i in the Wall of user n. Indeed $\left(\lambda^{(n)}+\mu^{(n)}\right)$ is the average number of posts per unit of time that enter the Wall of user n. Each of these posts replaces a post of origin i with probability $q_i^{(n)}$, due to the RANDOM eviction policy in Walls and the PASTA property. As a result, the left-hand side of equation (4.2) is the outgoing rate of posts of origin i from the Wall of user n. Obviously, $\mu^{(n)}p_i^{(n)}$ is the average number of posts of origin i per unit of time that arrive on the Wall of user n, due to the RANDOM selection policy in Newsfeeds.

We can re-write the set of equations for the Newsfeed and Wall, related to origin i, as a linear system in the form

$$\mathbf{p}_i = \mathbf{A} \cdot \mathbf{p}_i + \mathbf{b}_i \tag{4.3}$$

$$\mathbf{q}_i = \mathbf{C} \cdot \mathbf{p}_i + \mathbf{d}_i. \tag{4.4}$$

It is interesting to observe that the matrices $\mathbf{A}$ and $\mathbf{C}$ are the same for all origins i and only the vectors $\mathbf{b}_i$ and $\mathbf{d}_i$ differ. In fact it can be shown that the matrix $\mathbf{A}$ is sub-stochastic under very mild assumptions, thus the system has a unique solution. Given that the matrix inversion is computationally expensive for matrices of real-world size (order of hundreds of millions of users), we have further developed an iterative algorithm to find the solution, that makes use of the sparsity of $\mathbf{A}$.

Having solved for the influence of user i on any other user $n = 1, \ldots, N$ inside the platform, we would like to have a measure of global influence. We suggest the average influence, called here the Ψ-score, defined as follows,

$$\Psi_i = \frac{1}{N-1} \sum_{n \neq i} q_i^{(n)}, \tag{4.5}$$

but other options are also possible. Using the Ψ-score we assign a unique influence measure to each user i, so we can actually rank the users in the platform based on their influence. This is somewhat reminiscent of the PageRank score of web-pages, because it is based on a Markovian model and an importance score, but it is obviously very different, and specific of the OSP mechanism. To evaluate the quality of our ranking based on our new model and its Ψ-score, we have used real data traces from Twitter and Weibo. We show here only the results from the (smaller) Twitter dataset. It contains $2M$ posts from $180K$ users. For the comparison, in addition to our ranking based on the Ψ^{model}, we produce four more user rankings:

(a) We first derive an empirical user ranking based on the trace. Specifically, we find the empirical influence of a user i on a user n, defined as the average amount of time that a post from user i occupies the first slot in the Wall of user n, and we get the average empirical Ψ-score Ψ^{emu}. (b) We rank users based on their activity rate $\lambda^{(i)}$. (c) Another ranking comes from the number of followers (is graph-specific only). And (d) the final ranking comes from PageRank (also graph specific).

We show in Fig. 4.1 two types of plots: At the centre, a 2D scatter plot, where each point corresponds to a user and is the tuple of his predicted rank based on the empirical Ψ^{emu} (x-axis) and the model-based Ψ^{model} (y-axis). We observe a very satisfactory fit of the ranks produced by the model, since most points lie very close to the line $x = y$, which describes the ideal perfect match of ranks. The second type of plot illustrates a metric similar to the Jaccard similarity index to compare the two rankings, called *Common users proportion*. More precisely, if $\{u_1, \ldots, u_X\}$ are the top-X UserIDs according to the empirical list, whereas $\{v_1, \ldots, v_X\}$ the top-X UserIDs for the model list, we define the proportion of common users at depth X of the emulator list by

$$\mathcal{C}_X = \frac{|\{u_1, \ldots, u_X\} \cap \{v_1, \ldots, v_X\}|}{X}. \tag{4.6}$$

Note that this quantity converges to 1 as X grows to N, because the two full lists contain the same set of users. But for some given $X < N$ (e.g. top-10), the curve shows how well the model manages to rank users in relation to the emulator in the top-X positions. Fig. 4.1 (right) shows clearly that the model-based ranking explains almost 80% of the empirical ranking, much higher than graph-based rankings (PageRank and number of followers) or rankings based only on user activity.

We have thus developed a very interesting novel model that can explain sufficiently post diffusion in social platforms, by combining user position and activity in an appropriate way. The above results have been already published and a journal version is under review:

Publication [Giovanidis *et al.* 2019] A. Giovanidis, B. Baynat and A. Vendeville, "Performance Analysis of Online Social Platforms," IEEE INFOCOM 2019 - IEEE Conference on Computer Communications, 2019 (and submitted [journal-version] with Clémence Magnien).

4.3 Extensions

Based on the above modelling concept and results, I have obtained funding through an ANR French Young Researcher project. It provides financing for 4 years to conduct research on the subject of dynamic modelling of social platforms.

Funding (2020-2024) ANR JCJC 2019 - French Young Researcher Grant. Topic “FairEngine - Fair Online Social Platform Engineering”, 2 PhDs.

The project aims to extend the above model to more realistic platform mechanisms and user behaviour. Specifically, the aim is to include further aspects of the Newsfeed mechanism such as likes and comments. I want to propose Newsfeed curation algorithms that are fair to user content, and predict how a change in user activity or the Newsfeed impacts global influence. The project further aims to understand how orchestrated bot campaigns affect post diffusion and find protection mechanisms from the spread of malicious information. Another interesting goal of the project is to design advertising policies, and identify appropriate seeders to push advertising posts efficiently, without overwhelming the Newsfeed.

One of the two PhD positions from this project has been already covered.

• **Ph.D thesis supervision** of Ricardo José Lopez Dawn, “Stochastic Modeling and Data Analysis of Dynamic Post Diffusion in Online Social Platforms”, Co-supervised by Naceur Malouch. Sorbonne-LIP6, EDITE (start 2020)

The student currently works on the topic of user seeding for influence maximisation [Kempe *et al.* 2003], constrained on a monetary budget. We are using the dynamic activity model presented in the current section, something which brings several breakthroughs to an established NP-hard problem.

Chapter 5

Outro - the future

During the period covered in the HDR report (2012-today) I have focused on three main topics, namely (a) wireless access (and C-RAN architectures), (b) caching content at the wireless edge, and (c) modelling of online social platforms and information diffusion. The research was made possible through the collaboration and supervision of 4 PhD students (2 official) and several Masters theses. I have also taught systematically courses related to the mathematical tools used in the thesis, i.e. optimisation, queuing theory and machine learning.

In the future I plan to continue working on extensions of these topics, setting as principal priority to include more data analysis and machine learning tools:

• I will continue research on the social platform modelling through the ANR JCJC Jeune Chercheur project "Engineering Fair Online Social Platforms" (FairEngine), as described in the previous section. Here, i would like to include more diverse data and offer realistic solutions that scale well with the size of the problem, in the direction already taken in my GitHub page: [GitHub OSP].

• Concerning my research on telecommunication networks, I would like to give more emphasis on a data & machine learning (ML) approach. The rationale is that major advancements in telecom networks have emerged through physics and novel architectures, but after long development, technology has to a large extend matured. Current challenges are mostly related to the massive scale of data exchanged worldwide (see IoT and the autonomous driving application), rather than new hardware and design, with the potential exception of future quantum communications. Today, novelty is related to security, anomaly detection, prediction and control. To work on such topics, I have developed my own full semester course on "Data Analysis for Networks" taught at M2 level at Sorbonne-LIP6 (2019-2020 was first year), which contains the basic ML tools with real applications to telecom networks (28h course, 28h Python lab). This course is freely available on my GitHub page: [GitHub DataNets-Course] .

Furthermore, I have already first results in this direction, as presented in Ch. 2.6.1, where I apply a modified K-means clustering algorithms for C-RAN architectures, as well as in Ch. 3.7, where the problem of long-session recommendations is formulated as an MDP. The extension of this problem is by applying Reinforcement Learning tools to adapt the recommendations to actual user preferences. My further aim is to work on *federated learning* in the framework of Mobile Edge Computing, cellular communications and the internet.

Bibliography

[Abhari & Soraya 2009] A. Abhari and M. Soraya. *Workload generation for YouTube.* Multimedia Tools and Applications, vol. 46, no. 91, 2009. (Cited on page 29.)

[Akoum & Heath 2013] S. Akoum and R. W. Heath. *Interference Coordination: Random Clustering and Adaptive Limited Feedback.* IEEE Transactions on Signal Processing, vol. 61, no. 7, pages 1822–1834, 2013. (Cited on page 5.)

[Alfano *et al.* 2016] G. Alfano, M. Garetto and E. Leonardi. *Content-Centric Wireless Networks With Limited Buffers: When Mobility Hurts.* IEEE/ACM Transactions on Networking, vol. 24, no. 1, pages 299–311, 2016. (Cited on page 27.)

[Álvarez-Corrales 2017] Luis David Álvarez-Corrales. *Cooperative Communications in very large cellular Networks.* Theses, Telecom ParisTech, November 2017. (Cited on page 14.)

[Álvarez-Corrales *et al.* 2016] L. D. Álvarez-Corrales, A. Giovanidis and P. Martins. *Coverage Gains from the Static Cooperation of Mutually Nearest Neighbours.* In 2016 IEEE Global Communications Conference (GLOBECOM), pages 1–6. IEEE, 2016. (Cited on page 15.)

[Álvarez-Corrales *et al.* 2017] L. Álvarez-Corrales, A. Giovanidis, P. Martins and L. Decreusefond. *Wireless node cooperation with resource availability constraints.* In 2017 15th International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOpt), pages 1–8. IEEE, 2017. (Cited on page 19.)

[Andrews *et al.* 2011] J. G. Andrews, F. Baccelli and R. K. Ganti. *A Tractable Approach to Coverage and Rate in Cellular Networks.* IEEE Transactions on Communications, vol. 59, no. 11, pages 3122–3134, 2011. (Cited on pages 7, 8 and 11.)

[Andrews *et al.* 2014] J. G. Andrews, S. Singh, Q. Ye, X. Lin and H. S. Dhillon. *An overview of load balancing in hetnets: old myths and open problems.* IEEE Wireless Communications, vol. 21, no. 2, pages 18–25, 2014. (Cited on page 4.)

[Avranas & Giovanidis 2016] A. Avranas and A. Giovanidis. *Performance of spatial Multi-LRU caching under traffic with temporal locality.* In 2016 9th International Symposium on Turbo Codes and Iterative Information Processing (ISTC), pages 345–349. IEEE, 2016. (Cited on page 33.)

[Baccelli & Blaszczyszyn 2010] Francois Baccelli and Bartlomiej Blaszczyszyn. *Stochastic Geometry and Wireless Networks: Volume II Applications*. Foundations and Trends in Networking, vol. 4, no. 1?2, pages 1–312, 2010. (Cited on page 8.)

[Baccelli & Giovanidis 2013] F. Baccelli and A. Giovanidis. *Coverage by pairwise base station cooperation under adaptive geometric policies*. In 2013 Asilomar Conference on Signals, Systems and Computers, pages 748–753. IEEE, 2013. (Cited on page 11.)

[Baccelli & Giovanidis 2015] F. Baccelli and A. Giovanidis. *A Stochastic Geometry Framework for Analyzing Pairwise-Cooperative Cellular Networks*. IEEE Transactions on Wireless Communications, vol. 14, no. 2, pages 794–808, 2015. (Cited on page 11.)

[Baron *et al.* 2016] B. Baron, P. Spathis, H. Rivano and M. D. de Amorim. *Offloading Massive Data Onto Passenger Vehicles: Topology Simplification and Traffic Assignment*. IEEE/ACM Transactions on Networking, vol. 24, no. 6, pages 3248–3261, 2016. (Cited on pages 25 and 27.)

[Bastug *et al.* 2014] E. Bastug, M. Bennis and M. Debbah. *Living on the edge: The role of proactive caching in 5G wireless networks*. IEEE Communications Magazine, vol. 52, no. 8, pages 82–89, 2014. (Cited on page 26.)

[Bastug *et al.* 2015] E. Bastug, M. Bennis, M. Kountouris and M. Debbah. *Cache-enabled small cell networks: modeling and tradeoffs*. J Wireless Com Network, vol. 41, 2015. (Cited on page 25.)

[Bektas *et al.* 2008] Tolga Bektas, Jean-Francois Cordeau, Erhan Erkut and Gilbert Laporte. *Exact algorithms for the joint object placement and request routing problem in content distribution networks*. Computers & Operations Research, vol. 35, no. 12, pages 3860 – 3884, 2008. Part Special Issue: Telecommunications Network Engineering. (Cited on page 38.)

[Blaszczyszyn & Giovanidis 2015] B. Blaszczyszyn and A. Giovanidis. *Optimal geographic caching in cellular networks*. In 2015 IEEE International Conference on Communications (ICC), pages 3358–3363. IEEE, 2015. (Cited on page 31.)

[Borst *et al.* 2010] S. Borst, V. Gupta and A. Walid. *Distributed Caching Algorithms for Content Distribution Networks*. In 2010 Proceedings IEEE INFOCOM, pages 1–9, 2010. (Cited on page 25.)

[Böttger *et al.* 2018] Timm Böttger, Felix Cuadrado, Gareth Tyson, Ignacio Castro and Steve Uhlig. *Open Connect Everywhere: A Glimpse at the Internet Ecosystem through the Lens of the Netflix CDN*. SIGCOMM Comput. Commun. Rev., vol. 48, no. 1, page 28?34, April 2018. (Cited on page 24.)

[Breslau *et al.* 1999] L. Breslau, Pei Cao, Li Fan, G. Phillips and S. Shenker. *Web caching and Zipf-like distributions: evidence and implications.* In IEEE INFOCOM '99. Conference on Computer Communications. Proceedings. Eighteenth Annual Joint Conference of the IEEE Computer and Communications Societies. The Future is Now (Cat. No.99CH36320), volume 1, pages 126–134 vol.1, 1999. (Cited on page 28.)

[Brin & Page 1998] Sergey Brin and Lawrence Page. *The Anatomy of a Large-Scale Hypertextual Web Search Engine.* In Proceedings of the Seventh International Conference on World Wide Web 7, WWW7, pages 107–117, NLD, 1998. Elsevier Science Publishers B. V. (Cited on page 46.)

[Cha *et al.* 2010] M. Cha, H. Haddadi, F. Benevenuto and K. Gummadi. *Measuring user influence in twitter: The million follower fallacy.* In 4th International AAAI Conference on Weblogs and Social Media, ICWSM. AAAI, 2010. (Cited on page 46.)

[Checko *et al.* 2015] A. Checko, H. L. Christiansen, Y. Yan, L. Scolari, G. Kardaras, M. S. Berger and L. Dittmann. *Cloud RAN for Mobile Networks-A Technology Overview.* IEEE Communications Surveys Tutorials, vol. 17, no. 1, pages 405–426, 2015. (Cited on page 4.)

[Chen *et al.* 2018] Longbiao Chen, Dingqi Yang, Daqing Zhang, Cheng Wang, Jonathan Li and Thi-Mai-Trang Nguyen. *Deep mobile traffic forecast and complementary base station clustering for C-RAN optimization.* Journal of Network and Computer Applications, vol. 121, pages 59 – 69, 2018. (Cited on page 4.)

[China-Mobile 2011] China-Mobile. *White paper: C-RAN The Road Towards Green RAN.* Technical report Version 2.5, China Mobile Research Institute, October 2011. (Cited on page 4.)

[Choi *et al.* 2019] M. Choi, A. No, M. Ji and J. Kim. *Markov Decision Policies for Dynamic Video Delivery in Wireless Caching Networks.* IEEE Transactions on Wireless Communications, vol. 18, no. 12, pages 5705–5718, 2019. (Cited on page 41.)

[Cisco 2020] Cisco. *White paper: CISCO Annual Internet Report (update).* Technical report (2018-2023), Cisco public, March 2020. (Cited on pages 3 and 23.)

[Crovella & Bestavros 1997] M. E. Crovella and A. Bestavros. *Self-similarity in World Wide Web traffic: evidence and possible causes.* IEEE/ACM Transactions on Networking, vol. 5, no. 6, pages 835–846, 1997. (Cited on page 29.)

[Dan & Towsley 1990] Asit Dan and Don Towsley. *An Approximate Analysis of the LRU and FIFO Buffer Replacement Schemes.* SIGMETRICS Perform. Eval. Rev., vol. 18, no. 1, page 143?152, April 1990. (Cited on page 33.)

[DeGroot 1974] Morris H. DeGroot. *Reaching a Consensus.* Journal of the American Statistical Association, vol. 69, no. 345, pages 118–121, 1974. (Cited on page 46.)

[Dehghan *et al.* 2017] M. Dehghan, B. Jiang, A. Seetharam, T. He, T. Salonidis, J. Kurose, D. Towsley and R. Sitaraman. *On the Complexity of Optimal Request Routing and Content Caching in Heterogeneous Cache Networks.* IEEE/ACM Transactions on Networking, vol. 25, no. 3, pages 1635–1648, 2017. (Cited on pages 28 and 38.)

[Doppler *et al.* 2009] K. Doppler, M. Rinne, C. Wijting, C. B. Ribeiro and K. Hugl. *Device-to-device communication as an underlay to LTE-advanced networks.* IEEE Communications Magazine, vol. 47, no. 12, pages 42–49, 2009. (Cited on page 27.)

[Downey 2001] A. B. Downey. *The structural cause of file size distributions.* In MASCOTS 2001, Proceedings Ninth International Symposium on Modeling, Analysis and Simulation of Computer and Telecommunication Systems, pages 361–370, 2001. (Cited on page 29.)

[Erman *et al.* 2011] J. Erman, A. Gerber, M. Hajiaghayi, D. Pei, S. Sen and O. Spatscheck. *To Cache or Not to Cache: The 3G Case.* IEEE Internet Computing, vol. 15, no. 2, pages 27–34, 2011. (Cited on page 24.)

[Fagin 1977] Ronald Fagin. *Asymptotic miss ratios over independent references.* Journal of Computer and System Sciences, vol. 14, no. 2, pages 222 – 250, 1977. (Cited on page 33.)

[Fayazbakhsh *et al.* 2013] Seyed Kaveh Fayazbakhsh, Yin Lin, Amin Tootoonchian, Ali Ghodsi, Teemu Koponen, Bruce Maggs, K.C. Ng, Vyas Sekar and Scott Shenker. *Less Pain, Most of the Gain: Incrementally Deployable ICN.* SIGCOMM Comput. Commun. Rev., vol. 43, no. 4, page 147?158, August 2013. (Cited on page 24.)

[Fricker *et al.* 2012] C. Fricker, P. Robert and J. Roberts. *A versatile and accurate approximation for LRU cache performance.* In 2012 24th International Teletraffic Congress (ITC 24), pages 1–8, 2012. (Cited on page 33.)

[Garetto *et al.* 2016] Michele Garetto, Emilio Leonardi and Valentina Martina. *A Unified Approach to the Performance Analysis of Caching Systems.* ACM Trans. Model. Perform. Eval. Comput. Syst., vol. 1, no. 3, May 2016. (Cited on pages 26 and 33.)

[Gesbert *et al.* 2007] D. Gesbert, M. Kountouris, R. W. Heath, C. Chae and T. Salzer. *Shifting the MIMO Paradigm.* IEEE Signal Processing Magazine, vol. 24, no. 5, pages 36–46, 2007. (Cited on page 6.)

[Giannakas *et al.* 2020] T. Giannakas, A. Giovanidis and T. Spyropoulos. *MDP-based network friendly recommendations.* 2020. (Cited on page 41.)

[Giovanidis & Avranas 2016] A. Giovanidis and A. Avranas. *Spatial Multi-LRU Caching for Wireless Networks with Coverage Overlaps.* SIGMETRICS Perform. Eval. Rev., vol. 44, no. 1, page 403?405, June 2016. (Cited on page 33.)

[Giovanidis *et al.* 2012] A. Giovanidis, J. Krolikowski and S. Brueck. *A 0-1 program to form minimum cost clusters in the downlink of cooperating base stations.* In 2012 IEEE Wireless Communications and Networking Conference (WCNC), pages 940–945. IEEE, 2012. (Cited on page 9.)

[Giovanidis *et al.* 2015] A. Giovanidis, L. D. Álvarez Corrales and L. Decreusefond. *Analyzing interference from static cellular cooperation using the Nearest Neighbour Model.* In 2015 13th International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOpt), pages 576–583. IEEE, 2015. (Cited on page 15.)

[Giovanidis *et al.* 2019] A. Giovanidis, B. Baynat and A. Vendeville. *Performance Analysis of Online Social Platforms.* In IEEE INFOCOM 2019 - IEEE Conference on Computer Communications, pages 2413–2421, 2019. (Cited on page 50.)

[Giovanidis 2016] A. Giovanidis. *How to group wireless nodes together?* CoRR, vol. abs/1602.03906, 2016. (Cited on page 15.)

[Golrezaei *et al.* 2013] N. Golrezaei, A. F. Molisch, A. G. Dimakis and G. Caire. *Femtocaching and device-to-device collaboration: A new architecture for wireless video distribution.* IEEE Communications Magazine, vol. 51, no. 4, pages 142–149, 2013. (Cited on pages 25 and 27.)

[Häggström & Meester 1996] Olle Häggström and Ronald Meester. *Nearest neighbor and hard sphere models in continuum percolation.* Random Structures & Algorithms, vol. 9, no. 3, pages 295–315, 1996. (Cited on page 15.)

[Hao Che *et al.* 2002] Hao Che, Ye Tung and Zhijun Wang. *Hierarchical Web caching systems: modeling, design and experimental results.* IEEE Journal on Selected Areas in Communications, vol. 20, no. 7, pages 1305–1314, 2002. (Cited on page 33.)

[Holley & Liggett 1975] Richard A. Holley and Thomas M. Liggett. *Ergodic Theorems for Weakly Interacting Infinite Systems and the Voter Model.* Annals of Probability, vol. 3, no. 4, pages 643–663, 08 1975. (Cited on page 46.)

[Irmer *et al.* 2011] R. Irmer, H. Droste, P. Marsch, M. Grieger, G. Fettweis, S. Brueck, H. Mayer, L. Thiele and V. Jungnickel. *Coordinated multipoint: Concepts, performance, and field trial results.* IEEE Communications Magazine, vol. 49, no. 2, pages 102–111, 2011. (Cited on page 6.)

[Jaffrès-Runser & Jakllari 2018] K. Jaffrès-Runser and G. Jakllari. *PCach: The Case for Pre-Caching your Mobile Data.* In 2018 IEEE 43rd Conference on Local Computer Networks (LCN), pages 465–468, 2018. (Cited on page 25.)

[Jarray & Giovanidis 2016] C. Jarray and A. Giovanidis. *The effects of mobility on the hit performance of cached D2D networks.* In 2016 14th International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOpt), pages 1–8. IEEE, 2016. (Cited on page 36.)

[Jarray & Giovanidis 2018] C. Jarray and A. Giovanidis. *Successful file transmission in mobile D2D networks with caches.* Computer Networks, vol. 147, pages 162 – 179, 2018. (Cited on page 36.)

[Jungnickel *et al.* 2014] V. Jungnickel, K. Manolakis, W. Zirwas, B. Panzner, V. Braun, M. Lossow, M. Sternad, R. Apelfrojd and T. Svensson. *The role of small cells, coordinated multipoint, and massive MIMO in 5G.* IEEE Communications Magazine, vol. 52, no. 5, pages 44–51, 2014. (Cited on page 4.)

[Karakayali *et al.* 2006] M. K. Karakayali, G. J. Foschini and R. A. Valenzuela. *Network coordination for spectrally efficient communications in cellular systems.* IEEE Wireless Communications, vol. 13, no. 4, pages 56–61, 2006. (Cited on page 6.)

[Keeler *et al.* 2013] H. P. Keeler, B. Blaszczyszyn and M. K. Karray. *SINR-based k-coverage probability in cellular networks with arbitrary shadowing.* In 2013 IEEE International Symposium on Information Theory, pages 1167–1171, 2013. (Cited on page 31.)

[Kempe *et al.* 2003] David Kempe, Jon Kleinberg and Éva Tardos. *Maximizing the Spread of Influence through a Social Network.* In Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD ?03, page 137?146. Association for Computing Machinery (ACM), 2003. (Cited on pages 45 and 50.)

[Krioukov *et al.* 2010] Dmitri Krioukov, Fragkiskos Papadopoulos, Maksim Kitsak, Amin Vahdat and Marián Boguñá. *Hyperbolic geometry of complex networks.* Phys. Rev. E, vol. 82, page 036106, Sep 2010. (Cited on page 18.)

[Krolikowski *et al.* 2017] J. Krolikowski, A. Giovanidis and M. Di Renzo. *Fair distributed user-traffic association in cache equipped cellular networks.* In 2017 15th International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOpt), pages 1–6, 2017. (Cited on page 39.)

[Krolikowski *et al.* 2018a] J. Krolikowski, A. Giovanidis and M. Di Renzo. *A Decomposition Framework for Optimal Edge-Cache Leasing.* IEEE Journal on Selected Areas in Communications, vol. 36, no. 6, pages 1345–1359, 2018. (Cited on page 39.)

[Krolikowski *et al.* 2018b] J. Krolikowski, A. Giovanidis and M. Di Renzo. *Optimal Cache Leasing from a Mobile Network Operator to a Content Provider.* In IEEE INFOCOM 2018 - IEEE Conference on Computer Communications, pages 2744–2752, 2018. (Cited on page 39.)

[Krolikowski 2018] Jonatan Krolikowski. *Optimal Content Management and Dimensioning in Wireless Networks.* Theses, Université Paris-Saclay, November 2018. (Cited on page 38.)

[Kurose 2014] Jim Kurose. *Information-centric networking: The evolution from circuits to packets to content.* Computer Networks, vol. 66, pages 112 – 120, 2014. Leonard Kleinrock Tribute Issue: A Collection of Papers by his Students. (Cited on page 24.)

[Lee *et al.* 2013] K. Lee, J. Lee, Y. Yi, I. Rhee and S. Chong. *Mobile Data Offloading: How Much Can WiFi Deliver?* IEEE/ACM Transactions on Networking, vol. 21, no. 2, pages 536–550, 2013. (Cited on page 29.)

[Leonardi & Neglia 2018] E. Leonardi and G. Neglia. *Implicit Coordination of Caches in Small Cell Networks Under Unknown Popularity Profiles.* IEEE Journal on Selected Areas in Communications, vol. 36, no. 6, pages 1276–1285, 2018. (Cited on page 35.)

[Leonardi & Torrisi 2015] E. Leonardi and G. L. Torrisi. *Least recently used caches under the Shot Noise Model.* In 2015 IEEE Conference on Computer Communications (INFOCOM), pages 2281–2289, 2015. (Cited on page 29.)

[Li *et al.* 2015] J. Li, E. Björnson, T. Svensson, T. Eriksson and M. Debbah. *Joint Precoding and Load Balancing Optimization for Energy-Efficient Heterogeneous Networks.* IEEE Transactions on Wireless Communications, vol. 14, no. 10, pages 5810–5822, 2015. (Cited on page 9.)

[Lin *et al.* 2010] Y. Lin, L. Shao, Z. Zhu, Q. Wang and R. K. Sabhikhi. *Wireless network cloud: Architecture and system requirements.* IBM Journal of Research and Development, vol. 54, no. 1, pages 4:1–4:12, 2010. (Cited on page 4.)

[Newman 2005] MEJ Newman. *Power laws, Pareto distributions and Zipf's law.* Contemporary Physics, vol. 46, no. 5, pages 323–351, 2005. (Cited on page 28.)

[Newman 2010] M.E.J. Newman. Networks: An introduction. Oxford University Press, 2010. (Cited on page 46.)

[Nigam *et al.* 2014] G. Nigam, P. Minero and M. Haenggi. *Coordinated Multipoint Joint Transmission in Heterogeneous Networks.* IEEE Transactions on Communications, vol. 62, no. 11, pages 4134–4146, 2014. (Cited on pages 5 and 12.)

[Nygren *et al.* 2010] Erik Nygren, Ramesh K. Sitaraman and Jennifer Sun. *The Akamai Network: A Platform for High-Performance Internet Applications.* SIGOPS Oper. Syst. Rev., vol. 44, no. 3, pages 2–19, August 2010. (Cited on page 23.)

[Pan *et al.* 2018] C. Pan, M. Elkashlan, J. Wang, J. Yuan and L. Hanzo. *User-Centric C-RAN Architecture for Ultra-Dense 5G Networks: Challenges and Methodologies.* IEEE Communications Magazine, vol. 56, no. 6, pages 14–20, 2018. (Cited on pages 4 and 5.)

[Park *et al.* 2019] J. Park, S. Samarakoon, M. Bennis and M. Debbah. *Wireless Network Intelligence at the Edge.* Proceedings of the IEEE, vol. 107, no. 11, pages 2204–2239, November 2019. (Cited on page 43.)

[Paschos *et al.* 2016] G. Paschos, E. Bastug, I. Land, G. Caire and M. Debbah. *Wireless caching: technical misconceptions and business barriers.* IEEE Communications Magazine, vol. 54, no. 8, pages 16–22, 2016. (Cited on page 24.)

[Poularakis *et al.* 2014] K. Poularakis, G. Iosifidis and L. Tassiulas. *Approximation Algorithms for Mobile Data Caching in Small Cell Networks.* IEEE Transactions on Communications, vol. 62, no. 10, pages 3665–3677, 2014. (Cited on pages 28 and 38.)

[Roberts & Sbihi 2013] J. Roberts and N. Sbihi. *Exploring the memory-bandwidth tradeoff in an information-centric network.* In Proceedings of the 2013 25th International Teletraffic Congress (ITC), pages 1–9, 2013. (Cited on page 24.)

[Rosensweig *et al.* 2010] E. J. Rosensweig, J. Kurose and D. Towsley. *Approximate Models for General Cache Networks.* In 2010 Proceedings IEEE INFOCOM, pages 1–9, 2010. (Cited on page 25.)

[Sanguinetti *et al.* 2016] L. Sanguinetti, R. Couillet and M. Debbah. *Large System Analysis of Base Station Cooperation for Power Minimization.* IEEE Transactions on Wireless Communications, vol. 15, no. 8, pages 5480–5496, 2016. (Cited on page 9.)

[Seetharam *et al.* 2015] A. Seetharam, P. Dutta, V. Arya, J. Kurose, M. Chetlur and S. Kalyanaraman. *On Managing Quality of Experience of Multiple Video Streams in Wireless Networks.* IEEE Transactions on Mobile Computing, vol. 14, no. 3, pages 619–631, 2015. (Cited on page 25.)

[Shanmugam *et al.* 2013] K. Shanmugam, N. Golrezaei, A. G. Dimakis, A. F. Molisch and G. Caire. *FemtoCaching: Wireless Content Delivery Through Distributed Caching Helpers.* IEEE Transactions on Information Theory, vol. 59, no. 12, pages 8402–8413, 2013. (Cited on pages 24, 26, 28, 31 and 38.)

[Tanbourgi *et al.* 2014] R. Tanbourgi, S. Singh, J. G. Andrews and F. K. Jondral. *A Tractable Model for Noncoherent Joint-Transmission Base Station Cooperation.* IEEE Transactions on Wireless Communications, vol. 13, no. 9, pages 4959–4973, 2014. (Cited on pages 5 and 12.)

[Tatarinov *et al.* 1997] I. Tatarinov, A. Rousskov and V. Soloviev. *Static caching in Web servers.* In Proceedings of Sixth International Conference on Computer Communications and Networks, pages 410–417, 1997. (Cited on page 26.)

[Traverso *et al.* 2015] S. Traverso, M. Ahmed, M. Garetto, P. Giaccone, E. Leonardi and S. Niccolini. *Unravelling the Impact of Temporal and Geographical Locality in Content Caching Systems.* IEEE Transactions on Multimedia, vol. 17, no. 10, pages 1839–1854, Oct 2015. (Cited on page 29.)

[Tuholukova *et al.* 2017] A. Tuholukova, G. Neglia and T. Spyropoulos. *Optimal cache allocation for femto helpers with joint transmission capabilities.* In 2017 IEEE International Conference on Communications (ICC), pages 1–7, 2017. (Cited on page 28.)

[Woo *et al.* 2013] Shinae Woo, Eunyoung Jeong, Shinjo Park, Jongmin Lee, Sunghwan Ihm and KyoungSoo Park. *Comparison of Caching Strategies in Modern Cellular Backhaul Networks.* In Proceeding of the 11th Annual International Conference on Mobile Systems, Applications, and Services, MobiSys ?13, page 319?332. Association for Computing Machinery (ACM), 2013. (Cited on page 24.)

[Wu *et al.* 2015] J. Wu, Z. Zhang, Y. Hong and Y. Wen. *Cloud radio access network (C-RAN): a primer.* IEEE Network, vol. 29, no. 1, pages 35–41, 2015. (Cited on page 4.)

[Zhang *et al.* 2017] Haijun Zhang, Chunxiao Jiang, Mehdi Bennis, Merouane Debbah, Zhu Han and Victor C. M. Leung. *Heterogeneous Ultra-Dense Networks: Part 1.* Comm. Mag., vol. 55, no. 12, page 68?69, December 2017. (Cited on page 4.)